Analysis of speeches in the Spanish Parliament

Zhizhao Wang

A thesis submitted for the degree of Master in Artificial Intelligence

Facultat d'Informàtica de Barcelona

UNIVERSITAT POLITÈCNICA DE CATALUNYA

2018

Advisor: Prof. Lluís Padró Cirera

Co-Advisor: Prof. Enrique Romero Merino

Declaration

I, Zhizhao Wang, declare that this thesis titled, “What do politicians talk about? Analysis of speeches in the Spanish Parliament” and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a master degree at UPC.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at UPC or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Zhizhao Wang

2018

Acknowledgements

I would like to thank my thesis advisors Prof. Lluís Padró Cirera and Prof. Enrique Romero Merino of the Computer Science Department of UPC. They have helped a lot during my thesis work and have given me many advices about the algorithms and methods used in the research. Also, I want to express my thanks for giving me the access to the UPC computing cluster and helping me with the troubles I had while using it.

I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abstract

Political texts, such as the text obtained from parliamentary debates or electoral programs, are a valuable source of information. The words used by politicians in their discourses reveal their ideology, their preferences about public policy and the issues that they prioritize. This project will analyse the oral questions asked by all Spanish members of Parliament and try to find the correlation between the personal characteristics of politicians and their political preferences.

The first part of the project is a task of text classification. A transcription of the texts obtained from parliamentary debates and electoral programs in Spain from 1977 to 2017 have already been collected. Every text (topic) of the parliamentary debate or electoral program belongs to one Class or more, as well as one Subclass or more within the Class/Classes; this information has also been collected. Here by using different features and different classifiers, I tried to find the combination that could give the best result of classification.

The second part of the project is to find the correlation between politicians´ personal characteristics and their preference of the political topics. The brief introductions of the politicians of Legislature VI to Legislature XII have been collected. By using information extraction methods, the characteristics of the politicians can be presented in form of feature vectors. Here I used linear regression, conditional probability and mutual information to give quantitative information by correlating the text and the personal characteristics. This information obtained through text analysis will help address questions such as: Are certain kind of topics preferred/ignored by politicians with certain characteristic? Do personal characteristics of politicians (such as their gender or education) shape their political preferences?

Contents

1 Introduction

There is a close relationship between language and politics. Lakoff once said: “The politics is language, the language is power” [1]. The relationship between language and politics can be studied from two aspects: the political issues related to language and the language used in politics. The first one aims at studying the source of language, the relation of thinking and existence [2]. The second one concentrates on how the language complete the political purposes and functions [3]. This project, based on the political discourses they made, aims at finding the relation between the politicians themselves and their political preferences. Furthermore, by analysing their political preferences with their personal characteristics, the general influence of a certain characteristic or previous experience on the political preference could be known.

1.1 Political Discourse

Political discourse, or political text, is the language used in politics. McNair delimits the political discourse by the political communication. All the purposeful communication around politics can be considered as political communication, which includes:

1)The various forms of communication that politicians and other political participants engage in for a particular purpose;

2)The communication between politicians and non-politicians such as electors and reporters;

3) The news reports, editorials, and other media discussing politics on these politicians and their activities.

He says that these are all considered to be political discourse [4]. However, Wilson only accepts the politicians´ own speech as political discourse [5].

The political language has three main characteristics:

1) Political language has a strong purpose [3];
2) Political language should have clear participants [4];
3) Political language has diverse presented forms [4].

The political discourse not only makes the politicians to achieve their purposes, but also helps the public to understand better the politicians´ real thoughts.

1.2 Database Description

The database used during this project was provided by professor Aina Gallego of the Institut de Barcelona d'Estudis Internacionals and a Research Associate at the Institute of Political Economy and Governance. In the first part of this project, the collection of the oral questions of committee during 1977 and 2017 (Appendix 1) was used for text classification. In this dataset, every piece of information was constructed by ID, DATE, YEAR, MONTH, DAY, LEGISLATURE, TITLE, AUTHOR, GENDER, PARTY, PARLIAMENTARY GROUP, RESULT, TYPE, COMMITTEE NAME, COMMITTEE, CODE, SUBCODE, CODE_2, SUBCODE_2, Autor2, Autor3, Autor4 and número autors (see Fig 1).

ID	DATE	YEAR	MONTH	DAY	LEGISLATURE	TITLE	AUTHOR	GENDER	PARTY	PARLIAMENTA	RESULT	TYPE	COMMITTEE NAME	COMMITTEE (	CODE	SUBCODE
1	8/9/77	1977	8	9	0	Existencia de paraísos fiscales en las ciudades de Madrid, Barcelona y Bilbao,	Moreno Díez, Eduardo	1	GUCD	10	Convertido	Ordinaria	Comisión de Economía y Hacienda	3	1	107
2	8/9/77	1977	8	9	0	Posibilidad de elevar el mínimo exento del impuesto sobre el Patrimonio a dic	Muñoz Peirats, Joaquín	1	GUCD	10	Convertido	Ordinaria	Comisión de Economía y Hacienda	3	4	402
3	11/15/77	1977	11	15	0	Acceso a la Universidad de los alumnos de Medicina aprobados en Selectividad	Roca i Junyent, Miquel	1	GMC	4	Tramitado p	Ordinaria	Comisión de Educación	6	6	601
4	1/25/78	1978	1	25	0	Aplicación del Real Decreto 2499/1976, de 15 de octubre y Orden Ministerial	Colino Salamanca, Juan Luis	1	GS	2	Caducado	Ordinaria	Comisión de Agricultura	12	21	2104
5	2/9/78	1978	2	9	0	Política agrícola en Cuenca. (181/000013)	Zapatero Gómez, Virgilio	1	GS	2	Tramitado p	Ordinaria	Comisión de Agricultura	12	4	402
6	2/9/78	1978	2	9	0	Hospital de Motilla del Palancar (Cuenca). (181/000094)	Zapatero Gómez, Virgilio	1	GS	2	Tramitado p	Ordinaria	Comisión de Sanidad y Seguridad Social	10	3	322
7	2/20/78	1978	2	20	0	Personal laboral de las Juntas de Puertos. (181/000080)	Benítez Rufo, Manuel	1	GCO	1	Caducado	Ordinaria	Comisión de Trabajo	10	20	2004
8	2/23/78	1978	2	23	0	Retribuciones de los funcionarios y demás trabajadores de los Ayuntamientos,	Saavedra Acevedo, Jerónimo	1	GS	2	Convertido	Ordinaria	Comisión de Interior	4	20	2001
9	2/24/78	1978	2	24	0	Aplicación del Real Decreto 2499/1976, de 15 de octubre y Orden Ministerial	Colino Salamanca, Juan Luis	1	GS	2	Tramitado p	Ordinaria	Comisión de Agricultura	12	21	2104
10	2/24/78	1978	2	24	0	Política forestal del Ministerio y, de manera especial, sobre las causas de no a	Zapatero Gómez, Virgilio	1	GS	2	Tramitado p	Ordinaria	Comisión de Agricultura	12	21	2103
11	2/24/78	1978	2	24	0	Política de ordenación de cultivos y de manera especial del cultivo de la cebc	Fernández-Montesinos García,	1	GS	2	Tramitado p	Ordinaria	Comisión de Agricultura	12	4	402
12	2/24/78	1978	2	24	0	Política de ordenación de cultivos y de manera especial del cultivo del tomat	Bordes Vila, José Antonio	1	GS	2	Tramitado p	Ordinaria	Comisión de Agricultura	12	4	402
13	2/24/78	1978	2	24	0	Política en materia de fertilizantes. (181/000008)	Colino Salamanca, Juan Luis	1	GS	2	Tramitado p	Ordinaria	Comisión de Agricultura	12	4	402
14	3/1/78	1978	3	1	0	Fondo de Garantía Salarial (FGS). (181/000104)	Martín Toval, Eduardo	1	GSC	2	Caducado	Ordinaria	Comisión de Trabajo	10	5	510
15	3/2/78	1978	3	2	0	Cumplimiento general de ayudas al estudio de la Administración Institucional	Izquierdo Rojo, María	0	GS	2	Tramitado p	Ordinaria	Comisión de Presidencia	9	20	2004
16	3/3/78	1978	3	3	0	Creación en la Universidad de Murcia de una Facultad de Ciencias Económicas	Vivas Palazón, Francisco	1	GS	2	Tramitado p	Ordinaria	Comisión de Educación	6	6	601
17	3/9/78	1978	3	9	0	Política sobre los tipos de interés del Banco de Crédito Agrícola. (181/000064	Pin Arboledas, José Ramón	1	GUCD	10	Convertido	Ordinaria	Comisión de Hacienda	3	15	1501
18	3/9/78	1978	3	9	0	Explotación del puente José de Carranza por la Sociedad Bética de Autopista,	Sánchez Blanco, Jerónimo	1	GS	2	Tramitado p	Ordinaria	Comisión de Obras Públicas y Urbanismo	8	10	1002
19	3/16/78	1978	3	16	0	Declaración de zona de acción especial al municipio de Cervera del Río Alhar	Sáenz Cosculluela, Javier Luis	1	GS	2	Tramitado p	Ordinaria	Comisión de Interior	4	20	2001
20	3/16/78	1978	3	16	0	Alumbramiento de aguas subterráneas en la provincia de Almería. (181/00008	Gómez Angulo, Juan Antonio	1	GUCD	10	Decaído	Ordinaria	Comisión de Obras Públicas y Urbanismo	8	21	2104
21	3/17/78	1978	3	17	0	Decretos 1336/1977 y 320/1978 sobre Cámaras Agrarias, la gestación, democr	Zapatero Gómez, Virgilio	1	GS	2	Tramitado p	Ordinaria	Comisión de Agricultura	12	4	402
22	3/17/78	1978	3	17	0	Profesores de enseñanza permanente de adultos. (181/000054)	Gutiérrez Pascual, Vicente	1	GS	2	Tramitado p	Ordinaria	Comisión de Educación	6	6	604
23	3/17/78	1978	3	17	0	Propósitos del Gobierno en relación con la empresa privada Minas de Figaredo	Palacio Alvarez, Manuel	1	GS	2	Tramitado p	Ordinaria	Comisión de Industria y Energía	1	8	805
24	3/17/78	1978	3	17	0	Pérdidas originadas por varias avenidas en el río Ebro. (181/000083)	Cristóbal Montes, Angel	1	GS	2	Decaído	Ordinaria	Comisión de Obras Públicas y Urbanismo	8	21	2104
25	3/17/78	1978	3	17	0	Personal de centralitas dependientes de la Compañía Telefónica Nacional de E	Sáenz Cosculluela, Javier Luis	1	GS	2	Caducado	Ordinaria	Comisión de Trabajo	10	17	1706
26	3/23/78	1978	3	23	0	Trasvase Tajo-Segura. (181/000017)	Fuente y de la Fuente, Licinio	1	GAP	1	Convertido	Ordinaria	Comisión de Agricultura	12	21	2104
27	3/27/78	1978	3	27	0	Política de producción y comercialización del vino. (181/000011)	Sáenz Cosculluela, Javier Luis	1	GS	2	Tramitado p	Ordinaria	Comisión de Agricultura	12	4	404
28	3/28/78	1978	3	28	0	Política del Ministerio de Agricultura referente al paro agrícola. (181/000012)	Fuente y de la Fuente, Licinio	1	GAP	1	Convertido	Ordinaria	Comisión de Agricultura	12	4	402
29	3/28/78	1978	3	28	0	Política de orientación de las producciones agrarias. (181/000015)	Fuente y de la Fuente, Licinio	1	GAP	1	Convertido	Ordinaria	Comisión de Agricultura	12	4	402

Figure 1: Example of the dataset of oral questions.

Here, ID just indicates the position of a certain oral question in this dataset, which does not contain any information. The maximum number of ID is 25329, which means there are 25329 oral questions in this dataset all together. The name of the AUTHOR is given in the form of [family-name, first-name]. In Gender, ´1´ represents that the author is male and ´0´ for female. CODE and SUBCODE are the ID number of class and subclass that this certain oral question belongs to, such as economy or labor, their significance can be found in Spanish Codebook (Appendix 2). In this dataset there are 21 different classes in CODE; for every class of CODE, there could be several subclasses (SUBCODE) depending on the certain class, altogether there are 255 subclasses. These two were used as label in the text classification. A special fact that need to be paid attention is that some titles can belong to more than one subclasses at the same time (see Fig 2) or do not belong to any subclass (see Fig 3), but this kind of cases happen rarely (211 in total, less than 1%) and independently from the class, so that in the real classification step can be ignored and would not affect the classification result. So, later in the experiment for titles of the first case, the label will be one of those subclasses it belongs to and this unique subclass for classify will be chosen at random; the titles of second cases will not be used.

17751	6/4/08	2008	6	4	9	Fecha prevista para modificar el reg	Méndez Monasterio, Lourdes	0	GP	1	Tramitado por completo s	Comisión de Educación, Polit	6	13	1302; 1308;

Figure 2: An example of the multi-subclass title.

17905	7/11/08	2008	7	11	9	Opinión del Gobierno acerca del alto	Álvarez-Arenas Cisneros, María	0	GP	1	Convertido		Comisión de Trabajo e Inmigr	10	5

Figure 3: An example of the non-subclass title.

The second part of the project is trying to find out the correlation between politician´s characteristics and their political preference. In this part, the personal information of the politicians need to be used. Their information can be found in the file “diputados” (Appendix 3). The structure is: num, id, name, att and source (see Fig 4).

num	id	name	att	source
1	1	Gervasio Martínez-Villaseñor García	Diputado en las Legislaturas Constituyente, I, IV y V y Senador en la III. Casado. Cuatro hijos. Maestro nacional. Licenciado en Derecho. Abogado. Funcionario. Vicepresidente del Grupo Parlamentario de UCD 1978-82. Pre	leg6
2	2	José Vicente Beviá Pastor	Diputado en las Legislaturas II, III, IV y V y Senador en la Constituyente y I. Casado. Tres hijas. Licenciado en Filosofía y Letras (Sección de Filología Clásica). Catedrático de Griego de INB. Profesor Universitario. Vicepresid	leg6
3	3	Luis Alberto Aguiriano Forniés	Diputado en las Legislaturas III, IV y V y Senador en la Constituyente y II. Casado. Un Hijo. Economista y Técnico en publicidad	leg6
4	4	Presentación Urán González	Diputada en la V Legislatura. Dos hijos. Administrativa	leg6
5	5	Luis Felipe Alcaraz Masats	Diputado en las Legislaturas I y V. Doctor en Filología Románica por Granada. Miembro del Consejo Federal de IU.	leg6
6	6	María Jesús Aramburu del Río	Un hijo. Licenciada en Filosofía y Letras (Filología Hispánica). Profesora. Miembro del Consejo Federal de IU. Miembro de la Permanente, Ejecutiva, Consejo Andaluz de IU-CA, Los Verdes. Responsable de Formación Teóri	leg6
7	7	Julio Villarrubia Mediavilla	Casado. Dos hijos. Licenciado en Derecho. Secretario Interventor de Administración Local. Abogado. Secretario General del PSOE de Palencia y miembro del Comité Federal. Concejal, Primer Teniente de Alcalde y Portavo	leg6
8	8	Eugenio Enrique Castillo Jaén	Diputado en la Legislatura V. Casado. Dos hijas. Licenciado en Farmacia y empresario, especialista en alimentación y ecología. Del Comité Ejecutivo Provincial desde 1992 y de la Junta Directiva Nacional del PP desde 199	leg6
9	9	José Madero Jarabo	Diputado en la Legislatura V. Casado. Ingeniero Agrónomo. Funcionario de la Administración Central del Estado. Diputado en las Cortes de Castilla-La Mancha 1987-91.	leg6
10	10	José Luis Rodríguez Zapatero	Diputado en las Legislatura III, IV y V. Casado. Dos hijas. Licenciado en Derecho. Abogado. Profesor de Derecho Político. Secretario General del PSOE de León. Miembro del Comité Federal.	leg6
11	11	María Amparo Valcarce García	Casada. Una hija. Licenciada en Geografía e Historia. Inspectora de Educación. Consejera Comarcal del Bierzo 1991-95. Concejala del Ayuntamiento de Fabero desde 1991.	leg6
12	12	Javier Ignacio García Gómez	Casado. Una hija. Mediador de seguros. Concejal del Ayuntamiento de Jaén. Diputado Provincial de Jaén. Secretario Provincial del PP en Jaén.	leg6
13	13	Luis de Torres Gómez	Diputado en las Legislaturas IV y V y Senador en la III. Casado. Ocho hijos. Maestro Nacional. Concejal. Alcalde de Andújar. Diputado Provincial. Diplomado CESEDEN. Miembro del Comité Ejecutivo Regional y de la Junta D	leg6

Figure 4: Example of dataset of politician´s brief introduction.

Here, there are totally 2791 pieces of information in this dataset; num indicates the position of the information of certain politician in this dataset; id indicates the position within every source, which corresponds to the LEGISLATURE of the dataset of oral questions (Appendix 1). Both num and id do not contain any information for the experiment. In this dataset, the term “name” is given in form of [first-name family-name] which differs from the form in the dataset of oral questions (Appendix 1). The term “att” contains the information of politicians.

In the second part of the project, these two dataset need to be correlated. I used the name of the author (AUTHOR and name respectively) as well as legislature (LEGISLATURE and source respectively) as the match indicators because: by only using the name may cause a lot false positive matches because there may be a lot of politicians with the same name but actually different people. To avoid this kind of situation, extra information need to be used for matching. Then, I selected term “source” but not the information of legislature obtained in term “att” to match with “LEGISLATURE” for two reasons:

1) As we can see in Figure 4, not all the politicians´ “att” term contain legislature information.
2) It´s more accurate to select “source” because what can be confirmed is that the given personal information must be true during the period of legislature of “source”, but may not be true during the period of legislature obtained in “att”. Because the legislature numbers obtained in “att” are smaller than in “source”, which means they are periods earlier. In this case, given personal information may differ at that time. An easy example is that one may not get married before, but married now. So we cannot know if the non-time-invariant information such as number of children, education is the same at that time or not. If not, those characteristics at this time cannot be used to find the correlation between the oral questions of that time. In order to get the most accurate results, I didn´t take the early-aged oral questions into account and used ´source´ as the only match indicator with “LEGISLATURE”.

As I mentioned above, only the oral questions between LEGISLATURE 6 and 12 could be matched because the “diputados” file (Appendix 3) only contains the politicians´ information of that period of time.

2 Basic Text Data Processing

In this section, a generic introduction of basic steps for text data processing are explained in detail. The processing is done based on the linguistic semantic elements. The smallest unit of the semantic element is a word. The main steps are: i) document pre-processing, ii) text representation, iii) feature extraction/selection, iv) training and testing, v) performance evaluation.

A. Document Pre-processing

In this step, the size of the input text reduces significantly; the form of the words will be normalized as well. It involves processes such as:

1) Eliminate the non-text part: this is applied mainly in texts where the source is web pages. In the database used in this project (Appendix 1), there is a code after every title text (See Fig 1) and is supposed to be removed;
2) Determination of the sentence boundary [6]: it is not needed in this database because the every title is a sentence itself;
3) Stop-words elimination [6] [7]: Stop-words are words that occur with high frequency in any text regardless of its class/classes. They do not contain any information and will not benefit to the classification. There are stop-words in any languages (for example, ´a´, ´the´, ´of´, etc. in English; ´la´, ´el´, ´a´, etc. in Spanish). In many natural language processing packages there are already been collected can be removed easily by functions.
4) Normalization of words [6] [8]: There could be many forms of one word in the real text due to:

   Uppercase: hijo <= HIJO
   Plurality: hijo <= hijos
   Tense and Model: poner <= puso, poner <= pon
   Gender: bonito <= bonita

   There are two main method of normalization, stemming and lemmatization. Lemma is the base form of all its inflectional forms and itself is a word. Lemmatization is the process to reduce words into their lemma. Stemming is the process of reducing words to their root by eliminating prefix and/or suffix [8]. Stem is not a word and can be the same for different lemmas. For example: the lemma of word ´pongo´ is ´poner´, the stem is ´pon´.
5) Counting: word frequency statistics is the basis for feature selection/extraction and weight calculation.

B. Text Representation

After data pre-processing, the text can be represented as a document vector. There are some strategies of the representation of the text vector:

1) Vector Space Model (VSM) [13]: in the 1960s, Salton G. and his team proposed vector space model. The basic idea is: characterize text as a point in a vector space which is made of features, the form is $(w_1, w_2, \ldots, w_i)$, w_i is the weight for

*i*th feature. The degree of correlation between two texts is represented as the similarity between two points in the space, which is normally calculated by Euclidean distance or the cosine of the angle of the vectors. This model is widely used in practice. Common used text classification algorithms such as Support Vector Machine (SVM), K-nearest neighbor (KNN) and Naïve Bayes (NB) are all based on VSM.

2) Standard Boolean Model [14]: it can be considered as a special case of VSM. The weights can only be 1 or 0 depending on whether the feature exist in the text or not. In many cases, the results of classification by applying Boolean model are no worse than by using frequency of features as weights. The decision tree method, association rules method and Boosting method are based on Boolean model.
3) Probabilistic Model [14] [15]: it attempts to estimate the probability that the user will find a particular document relevant. Documents are ranked by their odds of relevance, which is the ratio of the probability that the document is relevant to the probability that the document is not relevant to the query. This model operates recursively and needs to be given initial guess of parameters for iteration. It also requires some simplifying assumptions such as independence between features and documents which sometimes are unrealistic. This model can be very hard to build and its complexity grows quickly.

After selecting the model, the next step is to choose the features as well as calculating their weights.

The features could be: words (lemmas/stems), multi-word (n-gram) [6] [16] and words with gender (for some languages such as Spanish). Words are the simplest kind of features. Multi-word can get the correlation between words which sometimes contains more information for classification. For example, ´red light district´ as a whole obviously contributes more than ´red´, ´light´ and ´district´. Words with gender is adding the gender information to the words feature.

There are some methods of calculating the weights of features:

1) Boolean weight [14]: as mentioned before, it is the simplest way and every feature will be treated as the same. The weights can only be 1 or 0. The shortcoming is that it cannot reflect the importance of a feature for a particular class.
2) Term Frequency (TF) [10] [17]: the weights are the frequency that a particular feature has shown up in the text. As the frequency of every feature can be largely distinct, it is an important indication of of the text category. In practice, as the total number of features for every document is different, the term frequency need to be normalized.
3) Inverse Document Frequency (IDF) [10]: IDF indicates the general importance of a feature. When the IDF is small, it means that this feature occurs generally in every category of texts, which means it does not contribute to classification although its TF could be large. Stop-words are the features that have large value

of TF but small value of IDF. This method is reasonable in term of the features´ information concentration but ignores the frequency.

4) Term Frequency-Inverse Document Frequency (TF-IDF) [10]: this method combines TF and IDF and only give large weights to those features with large TF and IDF at the same time.

C. Feature Extraction/Selection for text data

After represented in form of vectors, in most cases the features used for creating the vector space are redundant. Features usually mean significant words, multi-words or frequently occurring phrases indicative of text class in the context of text classification [9]. The dimension of the vectors is the number of all the features that have shown up in the text database after pre-processing. For a database there could be more than 100.000 features; but for every single text, it usually only contains very few of them. So the input for the later classification will be a high-dimensional-sparse matrix, which will increase the time-complexity and spatial-complexity of the classification algorithms. Therefore, feature extraction/selection needs to be applied to gain a lower dimensional vector at a lowest loss of classification information.

Feature selection is the process of selecting the most efficient and important features from the original feature set to form a new one; feature extraction is typically to map the high-dimensional feature space (usually linear mapping) to a low-dimensional space.

There are many methods of feature selection, the basic idea is: calculate a certain statistical value for every feature and set a threshold T (term-goodness criterion), eliminate those features of which values are less/more than T, the remaining features are considered to be efficient and important. By choosing the different statistical values, there are methods: Document Frequency (DF) [10] [11], Mutual Information (MI) [11], Information Gain (IG) [11], χ^2 statistic (CHI) [11] and Term Strength (TS) [11], etc. For feature extraction, the most popular methods are Principal Component Analysis (PCA) [12] and TF-IDF [10].

D. Training and Testing

In this step, an appropriate machine learning algorithm is applied to train the classifier. Some algorithms will be introduced in the next chapter. Before the application of the machine learning algorithm, the first thing to be done is the division of the data into training set and test set. Training set is used to train the classifier while the test set aims at evaluating the capacity of the classifier after training. Both sets should maintain the data contribution of the original dataset as much as possible and should be mutually exclusive. Two common strategies to divide the data are Hold-Out method and Cross-Validation. [18]

The idea of Hold-Out method is directly dividing the dataset into training set and test set randomly. Test set is used to evaluate the generalization capacity of the classifier. This

method is simple, but the result on the test set could be significantly influenced by the test set itself as it is created randomly. In this method, not all the data is used to build the classifier, so it is not appropriate for tasks with small database. [18]

In cross-validation, original dataset is divided into k sub-sets that are similar in size and mutually exclusive. In each subset the data distribution of the original set is maintained as much as possible. This can be achieved by stratified sampling [19]. Every time, taking the union of the k-1 subsets as training set and the left subset as test set. Thus there are k training-test combinations and k times of training, the final result is the average of k results. In this method, all the data is used for both training and test, which is ideal for small database tasks. [18] In a special case that k equal to the number of training patterns is known as leave-one-out cross-validation. [20]

E. Performance Evaluation

In order to know if the classifier performs well on the test set, some measures of performance should be used in the evaluation. The commonly-accepted performance evaluation measures are mostly focus on the settings where the examples are assumed to be identically and independently distributed (IID) [21].

Class\Recognized	as Positive	as Negative
Positive	tp	fn
Negative	fp	tn

Table 1: A confusion matrix for binary classification

A confusion matrix records the examples correctly and incorrectly recognized for each class and is built for measures of the quality of classification [21]. Table 1 presents a confusion matrix for binary classification, where *tp, fp, fn* and *tn* are true positive, false positive, false negative and true negative counts respectively.

Based on the IID assumption, the most used empirical measure is the *accuracy*, which does not distinguish between the number of correct labels of different classes [21]:

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn}$$

In order to estimate the performance of a classifier on different classes, *sensitivity* and *specificity* are introduced for positive and negative class respectively:

$$sensitivity = \frac{tp}{tp + fn}$$

$$specificity = \frac{tn}{fp + tn}$$

The measures *sensitivity* and *specificity* can evaluate the performance of classifier on the imbalanced database which *accuracy* fails. An easy example of the failure is that in the database the majority of examples are labelled as *negative* and only a few of them are labelled as *positive* (which is common in practice); in this case with predicting all

the testing examples as *negative* the classifier can still get good result in *accuracy*. The 3 measures perform well in binary classification but may not be enough to evaluate the classifiers for multi-class. The idea of multi-class classification is as follows: within a set of classes there is a class of special interest (regards as *positive*); other classes are left as a new multi-class classification or binary classification (if the left classes are 2) [21]. The measures on that selected positive class are:

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn} = sensitivity$$

$$F - measure = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision * recall}$$

All three measures distinguish the correct classification of labels within different classes and concentrate on one class (positive examples). *Recall* indicates relation between the correctly classified examples (true positives) and its misclassified examples (false negative), while *precision* aims at true positives and example misclassified as positives (false positive). *F-measure* takes both *precision* and *recall* into account and measures the performance of a classifier in general; When $\beta > 1$, it favors *precision*, and *recall* otherwise. It balances the two when $\beta = 1$, in this case it is called *F1*. [21]

3 Technical Review

3.1 Text Classification

Automatic text classification, or simply text classification, refers to the process by which a computer attributes a given text to a pre-defined class or classes accurately and efficiently, it is an important part of many data management tasks. In this section, 3 algorithms for text classification which are used later in the experiments of this project are introduced.

A. Naïve Bayes

Naïve Bayes classifiers are based on *Bayes´* theorem with strong (naïve) independence assumptions between the features. *Bayes´* theorem says for events A and B:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where $P(B) \neq 0$;
$P(A|B)$ is a conditional probability: the likelihood of event A occurring given that B is true;
$P(B|A)$ is also a conditional probability: the likelihood of event B occurring given that A is true;
$P(A)$ and $P(B)$ are the probability of observing A and B independently of each other which is known as the marginal probability. [22]

The probability measures a "degree of belief". The *Bayes´* theorem relates the degree of belief in a proposition before and after accounting for evidence. For proposition A and evidence B:
$P(A)$ is the *prior*, the initial degree of belief in A;
$P(A|B)$ is the *posterior*, the degree of belief having accounted for B: if a new evident is observed, the degree of belief of that event A happens will renew taking account the support of the evidence, which represented as $\frac{P(B|A)}{P(B)}$. In case of text classification, the event will be "this text belongs to class *i*", the evidence is the features of this text, the task of classification will be a statistical process of maximizing a *posterior* probability $P(A|B)$ (MAP). The result given by the classifier is actually the class with the maximum *posterior* probability by taking account the evidence. The naïve independent assumption in text classification specifically assumes that all attributes (features) of the examples (texts) are independently of each other given the context of the class. This assumption is clearly false in most real-world tasks, however *Naïve Bayes* performs classification very well [23].

There are two different generative models commonly used, both of which are based on the *Naïve Bayes* assumption. In one model, a document is represented by a vector of binary attributes (features) indicating if a certain word occurs or not in the document.

The frequency a word occurs is not taking into account. During the calculation, the probability of all the attributes values are multiplied including the probability of non-occurrence for words that do not occur in this document. [23] This kind of model is called “Multi-variate Bernouli Model”. The other model specifies that a document is represented by the set of word occurrences from the document. The frequency of the word occurrence is captured and only the probability of the word that occur are multiplied. This model is called “Multinomial Model”. [23]

B. Random Forests

The method of Random Forests is proposed by Breiman [24] which is based on bagging [25]. Both bagging and boosting [26] of classification trees belongs to “ensemble learning” methods, which generate many classifiers and aggregate their results. A simple classification tree sets a threshold for a feature, if the value of the feature is above the threshold then it is classified into a class; otherwise into the other class. In boosting, successive trees give extra weight to point incorrectly predicted by the earlier predictors, a weighted vote is taken for prediction in the end [26]. When using bagging with trees, successive trees do not depend on the earlier trees, instead, each is independently constructed by bootstrap sampling of the dataset, a majority vote is taken for prediction [25]. Random forests add an additional layer of randomness to bagging, which is how the trees are constructed. In a random forest, each node is split using the best split among a subset of predictors randomly chosen at that node instead of the best split among all variables which is used in standard trees [27]. This gives better results compared to many other classifiers and is proved to be robust against overfitting [24].

The algorithm for random forests is as follow [27]:

1. Get *n* samples from original dataset by using bootstrap sampling method;
2. For each sample, randomly sample *m* predictors from all the predictors and choose the best split from among those variables, grow a classification tree using this split;
3. Predict new data by aggregating the predictions of the n trees buy using majority votes.

C. Support Vector Machine

Suppose that there are a set of points that belongs to two groups in a two-dimensional space: if these points are linearly separable, then exist at least one line in this space that can separate of these points into two groups. The set of lines can be represented in form:

$$w_1 x + w_2 y + b = 0$$

where (w_1, w_2) is the direction of the line; b is a constant. Thus, the class of a new point (x_1, y_1) can be determined by the positivity/negativity of the inequality: if $w_1x_1 + w_2y_1 + b > 0$, the point belongs to the positive class; if $w_1x_1 + w_2y_1 + b < 0$, the point belongs to the negative class. As there can be a lot of values of w_1, w_2 and b that are able to separate the points, there are many candidate lines (splits). The best split of them should have the maximum margin from both groups of dataset points, which ensures the confidence of the classification. So the process of finding the best split transforms into maximizing the sum of the distances (margin) from the split to the closest points of each groups. These closest points which are used to maximize margin and to determine the value of (w_1, w_2) are called *support vectors*. The distance between support vectors and the line should be as large as possible, therefore the line should be in the middle of the margin and the constant b is determined in this way. As mentioned above, the *support vectors* are only a few specific data points, the SVM classifier does not use all the data points (see Fig 5). [28]

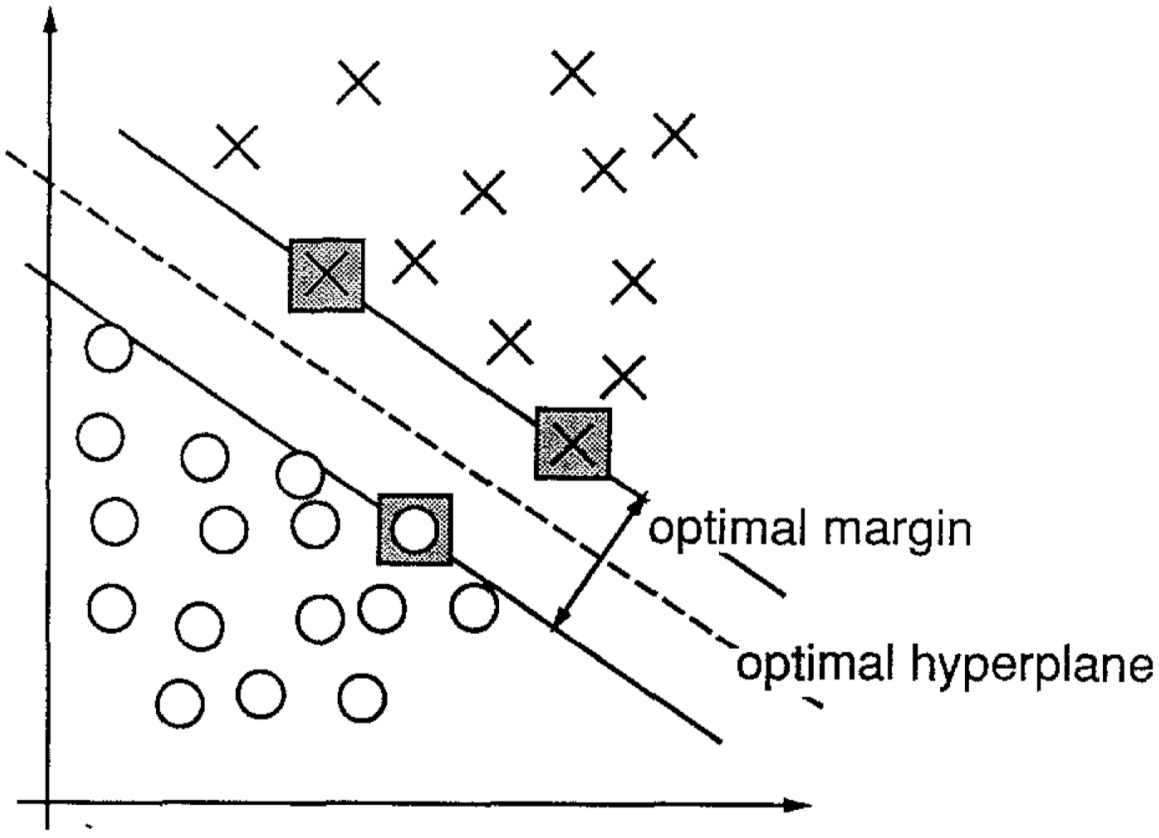

Figure 5: An example of a separable problem in a 2 dimensional space. The support vectors, marked with grey squares, define the margin of largest separation between the two classes [28].

In case of the separation in high dimensional space, the idea is the same: the task is to find a hyperplane that separates the points and has the maximum geometrical margin. For non-linearly separable data, in order to be able to split, a common solution is mapping the data into a higher dimensional space in which the data is linearly separable.

A problem is that the dimension of the space increases exponentially after mapping. The solution is using *kernel functions,* which are a set of functions to calculate the inner product (geometrical margin) of two vectors in the post-mapping space. *Kernel functions* can simplify the calculation of inner product in post-mapping space, which avoid the direct calculation in the higher dimensional space. Some common used kernel functions are: linear kernel, polynomial kernel, radial basis kernel (rbf), sigmoid kernel, etc. [29] Linear kernel is actually the kernel applied on the linearly separable data.

SVM classifiers are designed for binary classification problem. In order to solve multi-class problem, the common way is to combine the results of multiple binary SVM classifiers. There are two strategies:

1) One-versus-rest (OVR SVMs) [30]: in the training section, one particular class is regard as a class (positive) while all the rest are regarded as the other class (negative). Do this to every class, thus a number of binary SVM classifiers are obtained, the number of classifiers equals to the number of original classes. The classification of an unknown testing data will be determined by the classifier which has the highest value for this testing data among all the classifiers. The shortcoming of this strategy is that the training set may be highly biased because of the 1 vs rest. As all the rest of the classes are combines as negative, the examples of the negative class are obviously much more than the positive class examples.
2) One-versus-one (OVO SVMs/pairwise) [30]: in the training section, there is a classifier created for every pair of classes, so for a k-class problem, there are $\frac{k\,(k-1)}{2}$ classifiers altogether. The class of an unknown testing data will be the class that has the most votes by all the classifiers. The shortcoming is that the number of classifiers increases quickly as the increment of the number of classes.
3) Directed-acyclic-gragh (DAG-SVM) [30]: the training is the same as the one-versus-one strategy by solving $\frac{k\,(k-1)}{2}$ binary SVM classifiers. However, in the testing phase, it uses a rooted binary directed acyclic graph with $\frac{k\,(k-1)}{2}$ internal nodes and k leaves. Each node is a binary SVM classifier and the leave nodes are the classes; the test data starts at the root node and the binary function is evaluated. Then it moves either left or right depending on the output value (see Fig 6). In this way, the testing data goes through a path and finally reaches a leaf node which indicates the predicted class. The advantage of DAG-SVM is its generalization capacity [31] and faster testing time [30] than one-versus-one, however, they have the same testing time. The affecting of accumulative error from every layer of nodes is a big disadvantage.

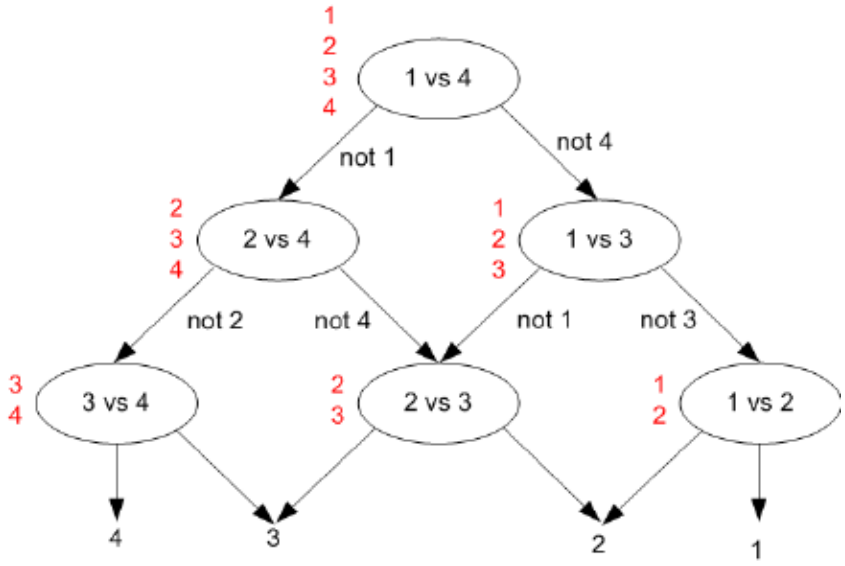

Figure 6: DAG-SVM for 4 classes[40].

3.2 Correlation Measures

The second part of the project is to confirm whether a particular characteristic of politicians correlates to their political preference; they are strong correlated or weak correlated and how the characteristic influences their political preference. Three methods of measuring correlation are introduced in this section.

A. Logistic Regression

Logistic regression is a generalized linear model usually used for classification. It belongs to statistical models and is usually taken to apply on dependent binary-class problems, however it can be generalized for dependent multi-value variable problems as well [32] [33]. The dependent variable value is often labelled as "0" or "1". The model is used to estimate the probability of the dependent variable based on a series of independent variables (features), where the dependent variable subjects to Bernouli distribution [34].

Dependent Variable	Probability
1	π
0	1- π

Table 2: Binary variable of Bernouli Distribution

Table 2 is an example of dependent variable which obeys Bernouli distribution, where π is the probability of "success" which satisfies: $0 < \pi < 1$. The definition of the odds ratio is:

$$Odds\ ratio = \frac{\pi}{1-\pi}$$

The probability π is represented by a logistic function (sigmoid function) [35], which takes real input and outputs a value between 0 and 1. Supposed that there are *n* independent features marked as x_i, the real input can be written as:

$$\beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

where β_0 is the bias and β_i is the coefficient of *i*-th feature.

Then the probabilities in form of logistic function are:

$$\pi = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1 - e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}$$

$$1 - \pi = \frac{1}{1 - e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}$$

The odds ratio can be written as:

$$Odds\ ratio = e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}$$

The logit form of odds ratio as:

$$\log(Odds\ ratio) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

In this form it can be seen obviously that the logit form of odds ratio is exactly a linear combination of features. The coefficients of features indicate the influence caused by every feature alone while the others are settled as constants:

if $\beta_i > 0$, then the *i*-th feature favors to the success of event (label "1");
if $\beta_i < 0$, then the *i*-th feature favors to the failure of event (label "0");
if $\beta_i = 0$, then the *i*-th feature does not correlate with the event.

Also, the absolute value of the coefficients shows the level of correlation between features and the dependent variable.

It is obvious that logistic regression can be used to analysis the correlation between every feature and the dependent variable (in our case is the class of political speech titles) by taking out the corresponding coefficient. However, the result totally depends on the trained model: which means that if the classification result is bad, the correlation results indicated by the coefficients are not truthful.

B. Conditional Probability

Suppose that there are two possible events A and B. Conditional probability is a measure of the probability of event A (the event of interest) given that event B has already occurred, usually written as $P(A|B)$ [36]. It is also called "posterior probability". Its definition is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

where $P(A \cap B)$ is the joint probability of two events, which indicates the probability that both events A and B occur; $P(B)$ is the marginal probability of event B, which indicates the probability that event B occurs.

The basic idea of measuring the level of correlation with conditional probability is by comparing the values of the conditional probabilities of the occurrence of different events of interest given the same event as evidence. Taking the case of our project as an example, event of interest A represents political category, event of evidence B represents the characteristic of politicians:

if $P(A = n|B = i) > P(A = m|B = i)$, the politicians with characteristic i seem to have the political preference on n than m.

However, there are some obvious problems of this method:

1) It cannot show if the two events are independent (non-correlated) with each other. In case of that A and B are independent, then $P(A|B) = P(A)$. $P(A)$ is the marginal probability of event A, whose value is between 0 and 1. So it is impossible to know if the correlation exists or not with only the value of conditional probability.
2) It cannot show that the two events are positively or negatively correlated. As shown above, this method depends on the pairwise comparisons to find out the preference relationship. However, the individual influence of a particular event of evidence to a particular event of interest cannot be revealed by the value of conditional probability. There could be cases like: $P(A = n|B = i) > P(A = m|B = i)$, but actually $B=i$ negatively influences to (correlates with) both $A=n$ and $A=m$.
3) Even the preference results obtained by the pairwise comparison could be wrong. Here is an example:

Political Speech Category	Characteristic: Married	Overall
Economy	350	1800
Human Rights	150	200

Table 3: Example of the failure by using conditional probability

As shown in the Table 3: all the married politicians have made 350 speeches about economy and 150 about human rights; while there are 1800 and 200 speeches about economy and human rights respectively raised by all the politicians.

Suppose that $P(Married) = \frac{350+150}{1800+200} = 0.25$, then we can calculate the probabilities:

$$P(Economy|Married) = \frac{\frac{350}{1800+200}}{P(Married)} = 0.7$$

$$P(Human\ Rights|Married) = 0.3$$

From the results we can see that it seems that married politicians have preference on economy than human rights; however we can see that $P(Economy) = 0.9$ and $P(Human\ Rights)$=0.1, so actually married politicians are morel likely to make speeches of human rights comparing with the overall. In this case, the conditional probability method gives a completely opposite answer because of ignoring the overall situation (prior of the event of interest).

To solve the three problems above, pointwise mutual information should be presented.

C. Pointwise Mutual Information (PMI)

In information theory, the mutual information (MI) of two random variables measures the mutual dependence between the two variables. It quantifies the average "amount of information" (unit: bit) obtained about one random variable by knowing the other one. The mutual information of two discrete random variables X and Y is defined as [37]:

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

where $p(x,y)$ is the joint probability of X and Y when $X = x$ and $Y = y$; $p(x)$ is the marginal probability of $X = x$ and $p(y)$ is the marginal probability of $Y = y$. Intuitively, it measures how much knowing one of these variables reduces uncertainty about the other. Mutual information has two properties [37]:

1) Non-negativity:

$$I(X;Y) \geq 0$$

2) Symmetry:

$$I(X;Y) = I(Y;X)$$

Mutual information is a measure of the correlation of two random variables, which means it refers to the average of all possible events. However, what we care about is the association of two particular events. In this case, pointwise mutual information (PMI) is used. It is defined as [38]:

$$pmi(x;y) = \log\frac{p(x,y)}{p(x)p(y)}$$

It can be seen that the mutual information is the expected value of pointwise mutual information. PMI could be negative or zero, but it still has the symmetric property.

Pointwise mutual information can be normalized (NPMI) between [-1,1] which 0 for independence, -1 and 1 for never occurring together and complete co-occurrence respectively [39]:

$$npmi = \frac{pmi(x;y)}{h(x,y)}$$

where $h(x,y) = -\log\ p(x,y)$. [37]

4 Experiments

In this section, the results of the classification of political speech titles by applying different classifiers and the correlation between politicians´ characteristics and their political preference are presented in 4.1 and 4.2 respectively. Some discussion and analysis are also given.

4.1 Classification of Political Speeches

In this part, the results obtained by three classifiers: Naïve Bayes Classifier (NB), Random Forest Classifier (RF) and Support Vector Machine with linear kernel (SVM) are given. In order to know the influence of some parameters, the results of the same classifier with different parameters are compared in form of line charts and tables. At the end the best results of every classifier and among all the classifiers are presented as well as the best result´s confusion matrix.

The data pre-processing is done by FreeLing, a natural language processing tool for Spanish as well as many other languages. The three potential kinds of features extracted to represent the dataset are a bag of lemmas (LM, with 20392 different lemmas), a bag of lemmas with gender (LG, size 11806) and a bag of lemma bi-grams (BLM, size 83510). These features are used to construct a vector that represents a single text in the dataset, in which the value of every feature can only be 1 or 0 (exist or not). This is because the texts are all titles which are short so that basically the maximum of the frequency of occurrence is 1 and it is more worthy the existence than the frequency. In every bag, there are some features that only occur once or twice in all the dataset texts which may not benefit to the classification. In order to know their influence, the eliminated features´ number of occurrence (EFN) is set as a parameter: *EFN=i* indicates that in this training session all the features that occur no more than *i* times are eliminated. Thus, the parameters for three classifiers in this experiment are:

NB: α (additive smoothing parameter), feature and EFN;
RF: n (number of trees in the forest), feature and EFN;
SVM: C (penalty parameter of the error term), feature and EFN;

where the possible values of feature are LM, LG, BLM and their combinations (LM+LG, LM+BLM, LG+BLM, LM+LG+BLM).

Basically, I considered the influence of every parameter to be independent from each other (although it may not be true sometimes) and found out the influence of a single parameter by controlling variables. In this way, a theoretical best result of a classifier can be obtained by using the combination of the best parameters. However, the actual best result is also given by searching diverse combinations of parameters. In the comparing part, the results shown are given in form of the accuracy of classifying CODE (Appendix 1), which is the F1-score of all classes in CODE. However, the precision, recall and F1-score for every class of the best result of three classifiers are given in detail at the end together with the confusion matrix.

There are 21 classes and 255 subclasses in the database. All the results obtained are based on a random split of 75% data for training and 25% for testing.

A. Naïve Bayes Classifier

In the experiment of this project, the classifier used is Bernouli model. The text examples are basically titles which are short. It is obvious that the appearance of a word is more worthy much more than its frequency of occurrence (basically the maximum will be 1).

1) Influence of α (feature = LM, EFN = 1):

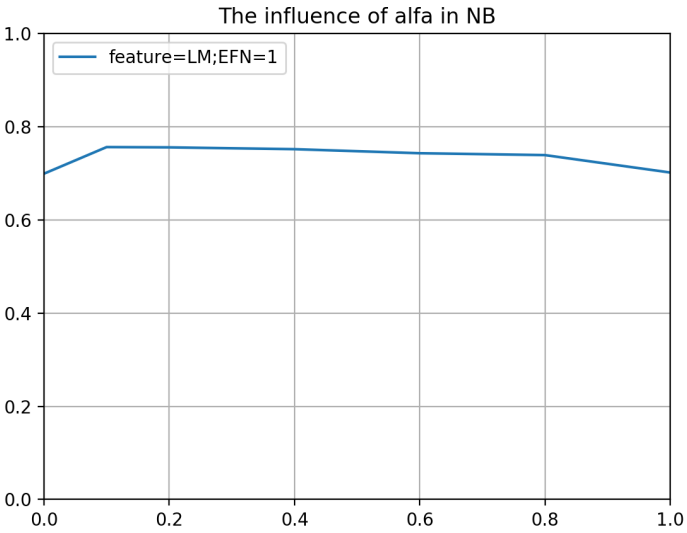

Figure 7: Accuracies of NB with different α

As shown in Figure 7, as the α gets bigger, the accuracy drops slightly, however it cannot be 0 (no smoothing). But in general, the smoothing parameter does not affect much to the results. The optimal value for α is 0.1 (slight smoothing).

2) Influence of the sets of features ($\alpha = 0.1$, EFN = 1):

Feature	Accuracy
LM	0.7558
LG	0.728
BLM	0.6891
LM+LG	0.7443
LM+BLM	0.7151
LG+BLM	0.7139
LM+LG+BLM	0.7206

Table 4: Accuracies of NB with different types of feature

As shown in Table 4, different types of feature indeed influence the result a lot. It can be seen that if multiple bags of feature are used together, the result is always better than the worst single-bag result and worse than the best single-bag result. For example: $Acc(BLM) < Acc(LG + BLM) < Acc\ (LG)$; $Acc(LG) < Acc(LM + LG) < Acc(LM)$. It can be understood in this way: for a multi-bag feature, only the better single-bag part is trying to give a good performance in classification, while the other is actually misleading the performance rather than giving extra information. There are altogether 25329 texts for 21 classes, while the number of features are 20392 for LM, 11806 for LG and 83510 for BLM. Since the number of features is quite a lot comparing with the number of examples, there could be possible correlations between different features which make NB classifier performs poorly. When different single-bag features are combined, this kind of possible correlations also occurs more frequently. Among all the single-bag feature, LM has the best performance, which is the optimal option for this parameter.

3) Influence of EFN (α = 0.1, feature = LM):

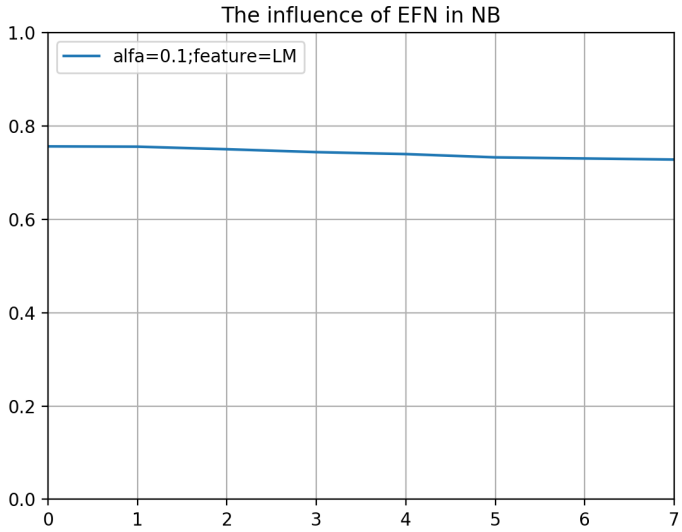

Figure 8: Accuracies of NB with different EFN

As shown in Figure 8, as the EFN gets bigger, more features are eliminated in the bag and the accuracy drops slightly. It is because although some features that occur with low frequency does not contain much information, it still has few benefit to the classifying, especially for any particular classes. But in general, EFN does not affect much to the results. The optimal value for EFN is 0 (without elimination).

4) Best of NB:
The best result in theory and in practice is the same: with the combination of α =0.1, feature = LM and EFN = 0. The accuracy is: 0.7564

B. Random Forest Classifier

1) Influence of n (feature = LM, EFN = 1):
   As shown in Figure 9, as n gets bigger, there are more trees in the forest, which means more classifiers are used in the classification, the result gets better very slightly. The tree number n must be greater than zero and there is no need to increase the number of trees endlessly as the result does not change a lot. Here I selected $n = 200$ to be the optimal.

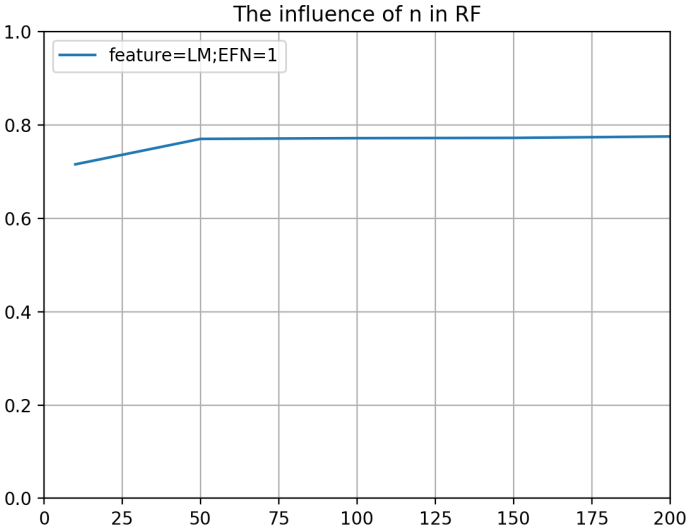

Figure 9: Accuracies of RF with different n

2) Influence of the sets of features ($n = 200$, EFN = 1):

Feature	Accuracy
LM	0.7738
LG	0.7732
BLM	0.719
LM+LG	0.7729
LM+BLM	0.7691
LG+BLM	0.7673
LM+LG+BLM	0.7703

Table 5: Accuracies of RF with different types of feature

From Table 5, it is shown that the influence of the type of the feature in RF is the same as in NB for the same reason. However, there is a special case is that the multi-bag feature LM+LG performs worse than both LM and LG. In this case, I think it is because single-bag feature LM and LG has similar accuracy, so the misleading information that one gives to each other affects both a little bit, which makes the

multi-bag feature performs slightly worse than both of the two. The optimal option is still LM.

3) Influence of EFN ($n = 200$, feature = LM):

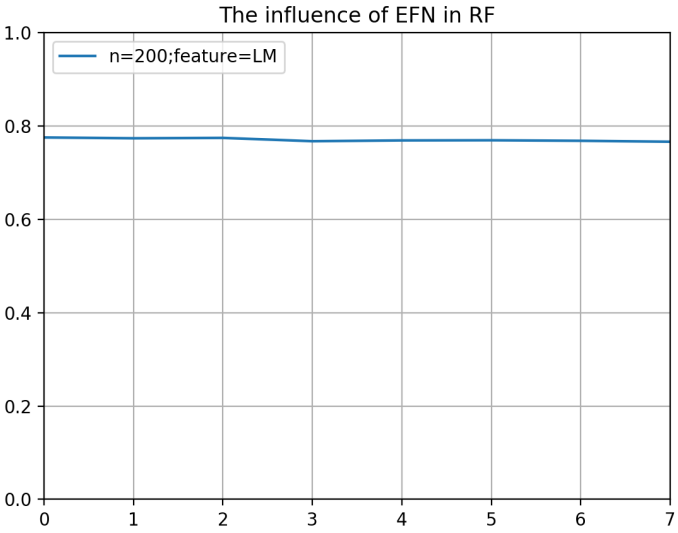

Figure 10: Accuracies of RF with different EFN

As shown in Figure 10, random forest classifier almost does not affect by EFN at all. The best result is given when none of the feature is eliminated, however the increasing of accuracy can be ignored. It proves that the rare features do not offer much information for the classification using RF classifier. As the EFN gets bigger, there is slight oscillation of the accuracies, it may be caused by the balancing of the loss of the useful information as well as the misleading information contained by the eliminated features. Anyway, EFN = 0 can be chosen as the optimal.

4) Best of RF:
Theoretical best result is given by combination of $n = 200$, feature = LM and EFN = 0, which is 0.7755. It is also the actual best result by using RF.

C. Support Vector Machine Classifier with linear kernel

As SVM classifiers are designed for binary classification, here I used the one-versus-rest (OVR) strategy to construct the classifier for this project. As the number of features is much more than the number of examples, SVM should be appropriate to deal with this project.

1) Influence of C (feature = LM, EFN = 1):

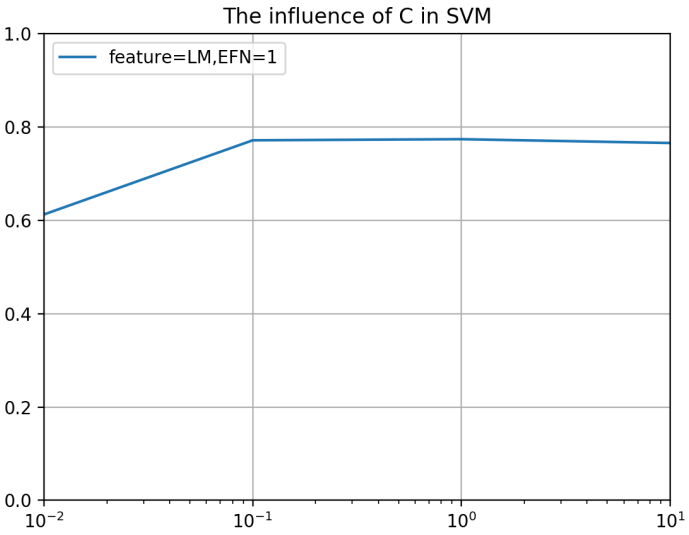

Figure 11: Accuracies of SVM with different C

As shown in Figure 11, SVM is relatively sensitive to the parameter C. The accuracy increases a lot from $C = 0.01$ to 0.1. However, the accuracy tends to be stable after C reaching 0.1. The best result is when $C = 0.1$ and 1, however the accuracy is really closed with the accuracy of $C = 10$.

2) Influence of the sets of features ($C = 0.1$, EFN $= 1$):

Feature	Accuracy
LM	0.7716
LG	0.7678
BLM	0.729
LM+LG	0.7825
LM+BLM	0.7812
LG+BLM	0.7828
LM+LG+BLM	0.7874

Table 6: Accuracies of SVM with different types of feature

From Table 6, we can see the behavior of SVM in term of the types of feature is quite different with NB and RF. In SVM, multi-bag feature always performs better than any of its constructing single-bag feature. Actually, the more features SVM uses, the better performance it gives. I think it is because SVM uses a different classifying method: SVM tries to find a hyperplane to separate the data points, and the hyperplane is only depending on some of the examples (support vectors), so its performance is not affected a lot by the increasing of examples or feature dimensions as it does not need to use all the data. So, it is relatively more robust. Also, if the support vectors can be described more in detail, the classifier will

perform better. Taking account of this, SVM should behave better with more features. Comparing with NB, SVM can deal with the possible correlations between features rather than assuming that the features are mutual independent (which rarely happens in practice). Because of this, SVM has the better capacity to find out the relation between features and examples as features´ increase and is less affected by the extra misleading information. Obviously, the optimal feature is the combination of the three bags.

3) Influence of EFN ($C = 0.1$, feature = LM):

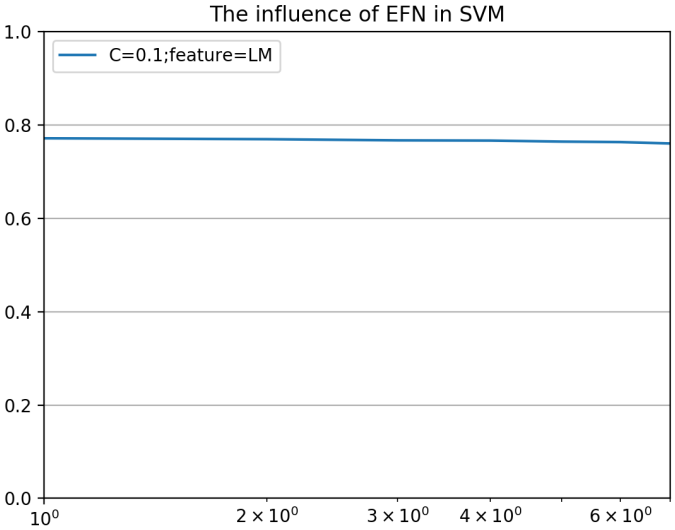

Figure 12: Accuracies of SVM with different EFN

As shown in Figure 12, SVM almost does not affect by EFN at all. The best result is given when none of the feature is eliminated, however the increasing of accuracy can be ignored. It proves that the rare features do not offer much information for the classification with SVM classifier. Like NB and RF, EFN = 0 is the best.

4) Best of SVM:
In theory, the best result of SVM occurs when $C = 0.1/1$, feature = LM+LG+BLM and EFN = 0; however actually the best result is 0.7874, which is obtained when $C = 0.1$, feature = LM+LG+BLM and EFN = 1.

Taking all the tree classifiers into account, the best result is given by support vector machine with linear kernel of which the parameters are: $C = 0.1$, feature = LM+LG+BLM and EFN = 1. The accuracy is 0.7874. The precisions, recalls and F1-scores for every class and subclass can be found in Appendix 4 together with the F1-score for all the classes, which is the accuracy on CODE.

Although in most cases it is shown that the results get better as the EFN decreases, I have to point out that it can dramatically save the RAM space needed for the experiments. The reason is that most of the features shown in the bags only occurs once or twice.

CODE	Precision	Recall	F1
1	0.6503	0.6611	0.6556
2	0.5548	0.5226	0.5382
3	0.8031	0.8093	0.8062
4	0.7638	0.8346	0.7976
5	0.7528	0.784	0.7681
6	0.7894	0.8366	0.8123
7	0.7099	0.6647	0.6866
8	0.874	0.75	0.8073
9	0.5926	0.5517	0.5714
10	0.8459	0.9232	0.8829
12	0.8463	0.8528	0.8496
13	0.6891	0.652	0.67
14	0.7353	0.6667	0.6993
15	0.7083	0.5129	0.595
16	0.8779	0.8367	0.8568
17	0.7802	0.6283	0.6961
18	0.56	0.3784	0.4516
19	0.75	0.8207	0.7838
20	0.6667	0.6526	0.6596
21	0.9056	0.8779	0.8915
23	0.6667	0.4167	0.5128

Table 7: Precision, recall and F1 of each class with SVM

Table 7 shows the precisions, recalls and F1-scores of each class (CODE) by using SVM classifier with linear kernel with its best parameters. It is a part of Appendix 4, the rest consists of the precision, recall and F1 of every subclass (SUBCODE) for every class. Particularly, the results of every subclass in Appendix 4 are obtained within of its class, which means the training and testing data for subclass classification are exactly from the data of subclasses´ class (true label). Therefore, if the classifier is used directly to classify the subclass, the performance (F1) should be the F1-score of the subclass in Appendix 4 multiplying the F1-score of its class in Table 7. For example, the F1-score of the classification of subclass 341 (tobacco) is 0.8062 (the F1-score of class 3) multiplying 0.8 (the F1-score of subclass 41 in class 3), which is 0.6445. The other detail is that in Appendix 4, there are some subclasses whose precision, recall and F1 are all set -1. This is due to the absence of the data of this subclass in the testing set. Both the

training and testing set hold the same data distribution as in the dataset, so it says that there are originally very few examples of this subclass in the dataset.

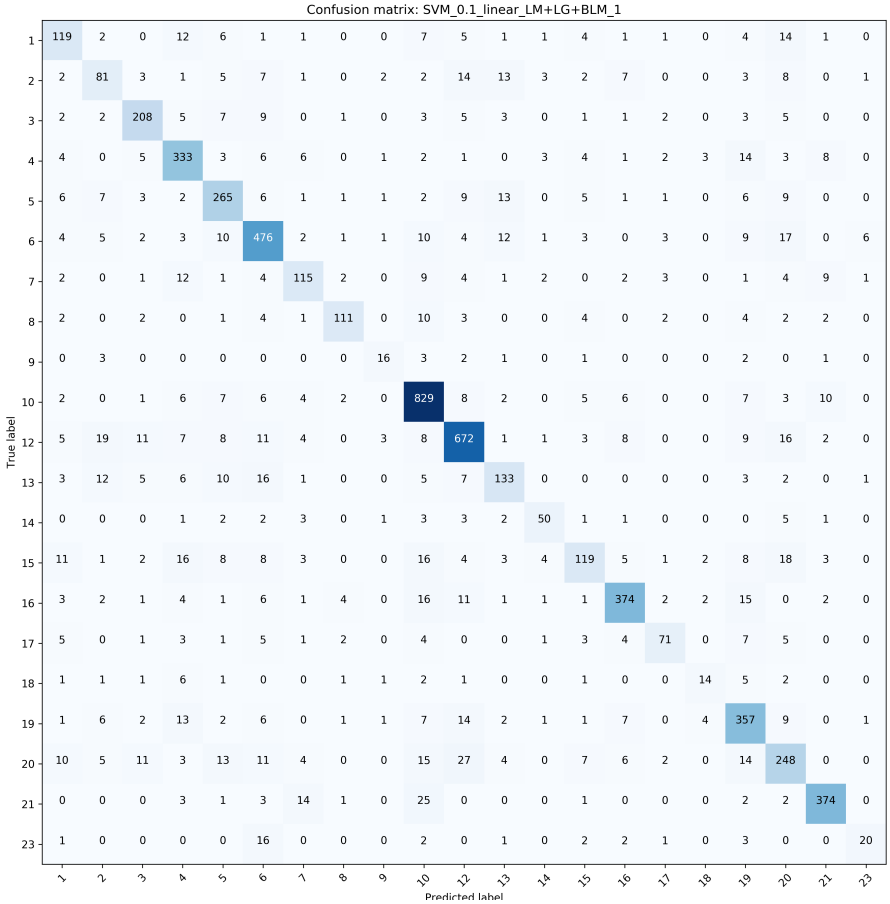

Figure 13: Confusion matrix of the classification with SVM

Figure 13 is the confusion matrix, in which the number is the real number of examples classified of the testing set.

From Table 7 we can see the F1-scores of class 2, 9, 15, 18 and 23 are relatively low. With the observation from Figure 13, we can figure out that these classes all have relatively lower number of examples (especially class 9) comparing with other classes. Because testing set has the same data distribution as in the training set, there are fewer examples of these classes used in training process and it leads to bad performance. However, it also has some relation with the speciality of the class itself: there are even fewer examples in class 8, but the classification of class 8 is even better than average. With the reference of Appendix 2, we know that class 8 is about energy, of which the political speeches often have the words (features) like “gas”, “nuclear”, “mina” or “electricidad” which are significantly differ from the words (features) of other classes.

An opposite case is class 23 (culture) with respect to the class 6 (education and culture), their words (features) have a lot of in common and it is hard to classify the examples of

class 23 from class 6. Also, we can see from the confusion matrix that the examples of class 18 (foreign trade) are mainly misclassified into class 4 (agriculture) and class 19 (foreign affairs). It is obvious that class 18 and class 19 will share a lot of features which are related with "foreign" such as "exterior", "internacional", "extrangero", "acuerdo", etc. In case of class 18 and class 4, it may be a little bit incomprehensible at first. However, some subclasses of class 4 such as 401 (agricultural trade), 402 (subsidies and agricultural regulation),404 (agricultural marketing and promotion) have a high similarity with some subclasses of class 18 such as 1802 (trade agreements, disputes and regulation) and 1803 (export promotion and regulation) in term of the features about trade, regulation and promotion.

Some concrete misclassified examples:
Title id 19126 "Grado de conocimiento de la lengua inglesa en el sistema universitario español por sectores (procedente de la pregunta con respuesta escrita al Gobierno número de expediente 184/15217). (181/001467)" of class 23. It is classified into class 6, I think it is because of the words "conocimiento" and especially "universitario", which are the typical words related with education.
Title id 4433 "Medidas para impedir la competencia del tomate marroquí al español en los mercados comunitarios. (181/001731)" of class 18. It is classified into class 4, in my opinion it is because of the words "tomate" and "mercado", which are the words related with agriculture and agricultural trade.
Title id 15497 "Objetivos y actuaciones del Convenio de Cooperación Internacional para la promoción transnacional del mercado turístico firmado entre España y Portugal. (181/001621)" of class 18. It is classified into class 19, I guess because of the words "cooperación", "internacional", "promoción", "transnacional" and "turístico". Actually class 18 and 19 are quite similar, but as class 19 has more examples for training and testing, the classifier tends to classify the uncertain example into class 19.

Both the number of examples and the classes themselves influence the performance of a classifier on each class. Some classes are highly similar in some aspects which make them difficult to be distinguished, but for this database it shows that the number of examples of each class plays a dominant role: in spite some classes have a lot in common, the ones with more examples always perform well. Although the titles of different classes may be similar, but there should be some specific features that can indicate its class in most cases. With more training examples, these specific features can be found more easily and accurately.

4.2 Correlation between Characteristics and Preferences

In this part, I used both a classification method and a statistical method to find out the correlation between politicians´ characteristic and their political preference. The classification method used is the logistic regression; the statistical method used is normalized pointwise mutual information (npmi). The result obtained by normalized pointwise mutual information is shown at the base 2 logarithm form.

The feature extraction of the politicians´ brief biography (Appendix 3) is done based on key words. The candidate features have been given by the database holder, which are shown in Table 8:

variable	**category**
género	hombre
género	mujer
estado_civil	casad
estado_civil	solter
estado_civil	divorciad
estado_civil	separad
estado_civil	viud
hijos	hij
hijos	sin_hij
partido	unidos_podemos
partido	cs
partido	en_comú_podem
partido	podemos_compromís_eupv
partido	a_la_valenciana
partido	erc_cat_sí
partido	cdc
partido	en_marea
partido	pnv
partido	eh_bildu
partido	cc_pnc
partido	podemos_iu_equo
partido	eaj_pnv
partido	asg
partido	cca_pnc
partido	erc
partido	bng
partido	partido_socialista_de_navarra

partido	psn
partido	popular
partido	psc
partido	psoe
partido	pp
partido	ciudadanos
partido	izquierda_unida
partido	convergència
partido	socialista
cargos	funcionari
cargos	portavoz_de_economía
cargos	portavoz_de_hacienda
cargos	autor
cargos	parlamentario_regional
cargos	alcalde
cargos	catedrático
cargos	profesor
cargos	parlamento_europeo
cargos	secretario_de_organización
cargos	delegad
cargos	portavoz
cargos	directiv
cargos	conseller
cargos	senador
cargos	ministr
cargos	diputada_provincial
cargos	diputado_provincial
cargos	empresario
cargos	técnic
educación	diplomad
educación	ingenier
educación	licenciad
educación	doctor
educación	máster
educación	bachiller
educación	estudiante
area_estudio	geografía
area_estudio	farmacia
area_estudio	historia
area_estudio	ciencias_políticas
area_estudio	médico

area_estudio	economista
area_estudio	ciencias_económicas
area_estudio	abogad
area_estudio	derecho
area_estudio	químicas
area_estudio	veterinaria
area_estudio	filosofía
area_estudio	sociología
area_estudio	empresariales
area_estudio	maestr
area_estudio	ciencias_biológicas
area_estudio	periodismo
area_estudio	banca
area_estudio	física

Table 8: Features of politicians´ characteristics

The key words of every feature are the possible variations of the feature or expressions that give the information of that the politician has this feature. For example: for feature "divorciad", the key words include "divorciado", "divorciada", "divorcio", "divorciar" and "divorció. In case of that one of the key words of a feature exists in the biography of a politician, the value of the corresponding feature in the vector of this politician is set to be 1; otherwise 0, which means this politician does not have this characteristic. In this way, the biography of each politician can be represented in form of a vector of which the dimension is the number of features in Table 8. The values in the vector are Boolean. In the experiment, there are some features of the Table 8 which are not owned by any politicians such as "en_marea" and "erc_cat_sí", in this case these features are eliminated to ensure the value of any feature in the vector is set to be 1 at least for one of the politicians. There are biographies of 2791 politicians altogether in Appendix 3.

The next step is to combine the politicians´ biography and their political preference into one single database. After linked with the CODE and SUBCODE of Appendix 1 by taking the politicians´ name (in Appendix 1: AUTHOR and in Appendix 3: name) and number of legislature (in Appendix 1: LEGISLATURE and in Appendix 3: source) as identifiers, a new database (Appendix 5) which holds the information of politicians´ characteristics and the classes and subclasses of the political speech titles they have presented is constructed. Here the number of legislature is also used as a identifier in order to avoid the problem of different politicians sharing the same name as well as ensure the timeliness of the biographies. After the linking, in Appendix 5 there are 5605 examples in total, which are 5605 different political speeches raised by less than 2791 politicians. Because some politicians did not propose any speeches according to the original databases.

Based on the new database (Appendix 5), we can begin to find out the correlation:

A. Classification method - Logistic Regression

In this method, CODE and SUBCODE are regarded as the labels respectively and the features representing the politicians´ characteristics are used for classifying. The coefficients of each feature for each class/subclass are extracted from the trained logistic regression classifiers. The coefficient for each class (CODE) is stored in Appendix 6; for subclass (SUBCODE), in Appendix 7. However, the results are not worth a trust. The accuracy of classification for CODE is 0.1176 and for SUBCODE is 0.0232, which are quite low. Therefore, although the coefficients can be calculated, they would not give any trustful information in this project.

The reasons for the bad performance of the classification are obvious:

1) There are 21 classes in CODE and 255 subclasses (SUBCODE) to be classified given only 5605 examples: it is a multi-class problem with small database.
2) Every politician can present speeches of different classes/subclasses, so in the database there could be a lot of situations that the same feature vector has many different labels: it is a multi-label problem.

It´s a multi-class multi-label classification at the same time, and the database is not big enough for the number of classes/subclasses either. It is logical that the performance is quite bad.

B. Statistical method – Normalized Pointwise Mutual Information

Statistical method is based on counting. The counts of the occurrence of every class/subclass as well as every feature are used later to calculate the statistical terms (probabilities). In the database (Appendix 5), the 5605 examples represent 5605 different political speech titles, so this number can be used to calculate the marginal probabilities of every political class of speech titles that all the politicians presented in general. However, this number does not have relation with the number of politicians so it cannot be used for the marginal probabilities of politicians´ having a particular characteristic. Because some politicians have presented more speeches and some have made less, the marginal probabilities of politicians´ having a particular characteristic here used is actually a weighted probability. It means that for the politicians who have presented more speeches, their characteristics have more weights comparing with the characteristics of the politicians who have presented less speeches. Thus, before the calculation of probabilities, two assumptions need to be proved.

Assumption 1: For this database, different speech titles raised by the same politician can be regarded as different speech titles of which each one is raised by a different politician and all these politicians have all the characteristics exactly the same.

Proof of Assumption 1:
Supposed that exist a politician who has the characteristics represented as a vector of features $\{f_1, \dots, f_i, \dots, f_N\}$, where f_i is the value of *i*-th feature and N is the number of features. This politician has presented M speeches of which CODE and SUBCODE are $\{C_1, \dots, C_j, \dots, C_M\}$ *and* $\{S_1, \dots, S_j, \dots, S_M\}$ where C_j and S_j are the class and subclass of j-th speech title. Therefore, the M titles of different speeches raised by this politician can be represented as:

$$\begin{pmatrix} C_1, S_1, f_1, \dots, f_i, \dots, f_N \\ \vdots \\ C_j, S_j, f_1, \dots, f_i, \dots, f_N \\ \vdots \\ C_M, S_M, f_1, \dots, f_i, \dots, f_N \end{pmatrix}$$

which is a M* (N+2) matrix, named as Matrix 1.
Supposed there a M different politicians, the CODE and SUBCODE of the speech presented by the *j*-th politician are marked as C_j and S_j. The *j*-th politician has the vector of feature as $\{f_{11}, \dots, f_{ij}, \dots, f_{Nj}\}$. Thus, the M titles of speeches raised by M different politicians can be represented as:

$$\begin{pmatrix} C_1, S_1, f_{11}, \dots, f_{i1}, \dots, f_{N1} \\ \vdots \\ C_j, S_j, f_{1j}, \dots, f_{ij}, \dots, f_{Nj} \\ \vdots \\ C_M, S_M, f_{1M}, \dots, f_{iM}, \dots, f_{NM} \end{pmatrix}$$

which is also a M* (N+2) matrix, named as Matrix 2.
Now with the knowledge of that all these M different politicians have exactly the same characteristics, which means:

$$f_{11} = f_{12} = \cdots = f_{1M} = f_1$$
$$\vdots$$
$$f_{i1} = f_{i2} = \cdots = f_{iM} = f_i$$
$$\vdots$$
$$f_{N1} = f_{N2} = \cdots = f_{NM} = f_N$$

By replacing the corresponding feature in the Matrix 2, Matrix 2 is exactly the same as Matrix 1. Assumption 1 is proved.

Assumption 2: For this database, those examples can be regarded as 5605 speeches which each one is presented by a different politician. However, different politicians may have all the characteristics the same.

Proof of Assumption 2:
By applying the assumption 1 on each of the politician, assumption 2 is proved.

With the assumption 2, all the necessary probabilities can be calculated:

$$p(c) = \frac{count(c)}{5605}$$
$$p(f) = \frac{count(f)}{5605}$$
$$p(c, f) = \frac{count(c, f)}{5605}$$

where: *c* is the class/subclass of the titles of speeches;

f is the feature of politicians;

count(c) is the number of titles of which the class/subclass is *c*;

count(f) is the number of politicians who has feature *f*;

count(c, f) is the number of titles which are presented by the politicians who have feature *f* and at the same time belongs to class/subclass *c*;

p(c) is the probability of titles belonging to class/subclass *c*;

p(f) is the probability of politicians having feature *f*;

p(c, f) is the probability of titles belonging to class/subclass *c* and at the same time raising by politicians who have feature *f*.

By counting the necessary data from the database, the normalized pointwise mutual information at base 2 logarithm is calculated and stored: Appendix 8 for CODE and Appendix 9 for SUBCODE. These results are trustful enough. In both, the values are limited in [-1, 1]. The more the value is closed to 0, the less correlation exists. The sign indicates the type of correlation. The blanks mean corresponding *count(c)* or *count(f)* is 0.

There is only one problem of this method, which also occurs by using any other methods: we have a table of numerical indicator between feature and class/subclass, but it is hard to find a threshold that we can say: if the value is bigger than the threshold then this feature is strongly correlated with this class/subclass. It is more like a stuff for domain experts. However, from the results I got, I could propose an intuitive threshold: as we all know, female politicians focus more on the problems of gender discrimination and rights than male politicians, this can be a golden standard. The subclass of “Gender discrimination and rights. Homosexuals discrimination and same-sex marriage” is 202, and the numerical indicator of subclass 202 and feature “mujer” is 0.164214462269531 according to Appendix 9. So 0.16 could be a “more than enough” value of threshold to judge the strong correlation on SUBCODE level: if the absolute value is equal or above 0.16, then we can say that the politicians who have this characteristic have preference on presenting or not presenting speeches of this class/subclass depends on the sign of the indicator. An interesting example is between subclass 202 and feature “hombre”, which the value of indicator is -0.1760077289331726. It reveals that male politicians even badly avoid to mention anything about gender discrimination, homosexual discrimination and same-sex marriage. Something interesting I found is that the politicians who were doctors or studied medicine are quite neutral about any topics

related with health (subclasses of class 3); however, those who were technician or engineer have preference talking about satellites and other space technology with commercial use (subclass 1704) and science technology transfer and scientific international cooperation (subclass 1705).

For CODE, the numerical indicator of class 2 and feature “mujer” can be considered as the threshold, which is 0.10856578722505628. With this as a more-than-enough standard, we can see some examples: the female politicians are pretty unlikely to present speeches about the defence and military (class 16); the politicians who studied economy prefer the topics of education and culture (class 6), energy (class 8) and especially foreign trade (class 18) while less prefer the topics of foreign affairs (class 19) and governmental issues (class 20). The politicians who studied geography and political science do not like the topic of technology and research (class 17). An interesting observation is that the professors (feature “catedrático”) do not prefer technology and research topic either. A quite logical case is that the politicians who studied banking prefer macro-economy (class 1), social policy (class 13), commerce and banking (class 15) topics.

For the features “casad”, “hij” and “sin_hij”, there is not any numerical indicator that reaches the threshold for neither CODE nor SUBCODE. It means the characteristics of married, with/without child do not influence the politicians´ political preference in general.

By applying the experimental thresholds mentioned above, the important/strong correlations are taken and displayed in Appendix 10 and Appendix 11, for CODE and SUBCODE respectively.

References

[1]. Lakoff, Robin Tolmach, “Talking Power: The Politics of Language in Our Lives” [M]. USA: BasicBooks,1990.

[2]. Holborow, Marnie, “The Politics of English” [M]. London: Sage Publications, 1999.

[3]. Chilton, P. & Schaffner, C. Discourse and Politics [A]. InDiscourseasSocial Interaction [C]. vanDijk. (ed) London: Sage Publications Ltd, 1997: 206-230

[4]. McNair, Brian, “An introduction to Political Communication” [M]. London: Routledge, 1999.

[5]. Wilson, John, “Politically Speaking: The pragmatic Analysis of Political Language” [M]. Oxford: Basil Blackwell, 1990

[6]. Zhang W., Yoshida T., and Tang X. “Text classification using multi-word features. ” In proceedings of the IEEE international conference on Systems, Man and Cybernetics, pp. 3519 – 3524, 2007

[7]. Kim S., Han K., Rim H., and Myaeng S. H. “Some effective techniques for naïve bayes text classification. ” IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457-1466, 2006

[8]. Porter M. F. “An algorithm for suffix stripping.” Program, 14 (3), pp. 130-137, 1980.

[9]. M. K. Dalal, M. A. Zaveri, “Automatic Text Classification: A Technical Review”, International Journal of Computer Applications (0975 8887), Volume 28 No.2, August 2011.

[10]. Jones K. S. “A statistical interpretation of term specificity and its application in retrieval.” Journal of Documentation, Vol. 28, No. 1, pp. 11-21, 1972.

[11]. Y. Yang and J. Pedersen. “A comparative study on feature selection in text categorization.” International Conference on Machine Learning (ICML), 1997.

[12]. Pearson, K. LIII. “On lines and planes of closest fit to systems of points in space.” Philosophical Magazine Series 6, 2, pp. 559-572, 1901.

[13]. G. Salton, A. Wong, and C. S. Yang. “A vector space model for automatic indexing.” Communications of the ACM (CACM), 18 (11):613–620, November 1975.

[14]. Arash Habibi Lashkari, Fereshteh Mahdavi, Vahid Ghomi, "A Boolean Model in Information Retrieval for Search Engines," in 2009 International Conference on Information Management and Engineering, pp. 385-389, 2009

[15]. R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. ACM Press, New York, 1999.

[16]. Church K. W., and Hanks P. “Word association norms, mutual information and lexicography.” Computational Linguistics, Vol. 16, No. 1, pp. 22-29, 1990.

[17]. Luhn, Hans Peter. "A Statistical Approach to Mechanized Encoding and Searching of Literary Information". IBM Journal of research and development. IBM. 1 (4): 315, 1957.

[18]. R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection”. in: Proceedings of the 14th International Conference on Artificial Intelligence (IJCAI), San Mateo, CA, Morgan Kaufmann, Los Altos, CA, pp. 1137–1143., 1995

[19]. Hunt, Neville; Tyrrell, Sidney. "Stratified Sampling". Webpage at Coventry University. Archived from the original on 13 October 2013. Retrieved 12 July 2012.

[20]. G.C. Cawley, N. Talbot. “Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers”. Pattern Recognition 36 (2003) pp.2585–2592, 2003.

[21]. M. Sokolova, N. Japkowicz, S. Szpakowicz. “Beyond accuracy, F-score and ROC: a family of

discriminant measures for performance evaluation". In: Australian conference on artificial intelligence, vol 4304. LNCS, pp 1015-1021, 2006.

[22]. Staurt, A. and Ord, K., "Kendalls Advanced Theory of Statistics. Volume 1: Distribution Theory". Edward Arnold, London, 1998.

[23]. Andrew McCallum and Kamal Nigam, "A Comparison of Event Models for Naive Bayes Text Classification". Workshop on Learning for Text Categorization, In the Fifteenth National Conference on Artificial Intelligence (AAAI), 1998.

[24]. L. Breiman, "Random forests". Machine Learning, 45 (1):5–32, 2001.

[25]. L. Breiman, "Bagging predictors". Machine Learning, 24 (2):123–140, 1996.

[26]. R. E. Shapire and Y. Singer, "BoosTexter: A System for Multi-Label Text Categorization". Draft, Mars 1998.

[27]. Liaw, A. & Wiener, M., "Classification and regression by randomForest". R News 2 (3): 18—22, 2002

[28]. C. Cortes and V. Vapnik, "Support-vector networks". Machine Learning, 20:273-279, November, 1995.

[29]. Hsu, C. W., Chang, C. C., & Lin, C. J., "A practical guide to support vector classification". Available at: http://www.csie.ntu.edu.tw/~cjlin/ papers/guide/guide.pdf., 2003.

[30]. Hsu, C. W. and Lin, C. J., "A comparison of methods for multi-class support vector machines". IEEE Trans. Neural Netw. 13, 2, 415–425, 2002.

[31]. J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAG's for multiclass classification," in Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, vol. 12, pp. 547–553, 2000.

[32]. Walker S.H. and Duncan D.B., "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54 (1/2): 167–178, 1967.

[33]. Cox D.R., "The regression analysis of binary sequences (with discussion)". J Roy Stat Soc B. 20 (2): 215–242, 1958.

[34]. James Victor Uspensky, "Introduction to Mathematical Probability". McGraw-Hill, New York, page 45, 1937.

[35]. Hosmer David W. and Lemeshow Stanley, "Applied Logistic Regression (2nd ed.)". Wiley, 2000.

[36]. Gut Allan, "Probability: A Graduate Course (Second ed.)". New York, NY: Springer, 2013.

[37]. Cover T. M. and Thomas J. A., "Elements of Information Theory". Wiley ed, 1991.

[38]. Kenneth Ward Church and Patrick Hanks, "Word association norms, mutual information, and lexicography". Comput. Linguist. 16 (1): 22–29, 1990.

[39]. Bouma, Gerlof, "Normalized (Pointwise) Mutual Information in Collocation Extraction". Proceedings of the Biennial GSCL Conference, 2009.

[40]. P. Cheol-Sun, and C. Jun-Ho, N. Sun-Phil, and J. Won, "Automatic Modulation Recognition of Digital Signals using Wavelet Features and SVM," in Proc. 10th ICACT, Phoenix Park, Korea, vol. 1, Feb. 2008, pp. 387-390.

Appendices

Here only a part of each appendix is shown because of the limitation of the space. The full version can be seen in the corresponding files.

1. oral-questions-committee-1977-2017 (first 30 examples):

ID	DATE	YEAR	MONTH	DAY	LEGISLATURE	TITLE	AUTHOR
1	8/9/77	1977	8	9	0	Existencia de paraísos fiscales en las ciudades de Madrid, Barcelona y Bilbao,	Moreno Díez, Eduardo
2	8/9/77	1977	8	9	0	Posibilidad de elevar el mínimo exento del impuesto sobre el Patrimonio a die	Muñoz Peirats, Joaquín
3	11/15/77	1977	11	15	0	Acceso a la Universidad de los alumnos de Medicina aprobados en Selectividad	Roca i Junyent, Miquel
4	1/25/78	1978	1	25	0	Aplicación del Real Decreto 2499/1976, de 15 de octubre y Orden Ministerial (	Colino Salamanca, Juan Luis
5	2/9/78	1978	2	9	0	Política agrícola en Cuenca. (181/000013)	Zapatero Gómez, Virgilio
6	2/9/78	1978	2	9	0	Hospital de Motilla del Palancar (Cuenca). (181/000094)	Zapatero Gómez, Virgilio
7	2/20/78	1978	2	20	0	Personal laboral de las Juntas de Puertos. (181/000080)	Benítez Rufo, Manuel
8	2/23/78	1978	2	23	0	Retribuciones de los funcionarios y demás trabajadores de los Ayuntamientos,	Saavedra Acevedo, Jerónimo
9	2/24/78	1978	2	24	0	Aplicación del Real Decreto 2499/1976, de 15 de octubre y Orden Ministerial (	Colino Salamanca, Juan Luis
10	2/24/78	1978	2	24	0	Política forestal del Ministerio y, de manera especial, sobre las causas de no a	Zapatero Gómez, Virgilio
11	2/24/78	1978	2	24	0	Política de ordenación de cultivos y de manera especial del cultivo de la cebo	Fernández-Montesinos García, A
12	2/24/78	1978	2	24	0	Política de ordenación de cultivos y de manera especial del cultivo del tomat	Bordes Vila, José Antonio
13	2/24/78	1978	2	24	0	Política en materia de fertilizantes. (181/000008)	Colino Salamanca, Juan Luis
14	3/1/78	1978	3	1	0	Fondo de Garantía Salarial (FGS). (181/000104)	Martín Toval, Eduardo
15	3/2/78	1978	3	2	0	Cumplimiento general de ayudas al estudio de la Administración Institucional	Izquierdo Rojo, María
16	3/3/78	1978	3	3	0	Creación en la Universidad de Murcia de una Facultad de Ciencias Económicas	Vivas Palazón, Francisco
17	3/9/78	1978	3	9	0	Política sobre los tipos de interés del Banco de Crédito Agrícola. (181/000064	Pin Arboledas, José Ramón
18	3/9/78	1978	3	9	0	Explotación del puente José de Carranza por la Sociedad Bética de Autopista,	Sánchez Blanco, Jerónimo
19	3/16/78	1978	3	16	0	Declaración de zona de acción especial al municipio de Cervera del Río Alhan	Sáenz Cosculluela, Javier Luis
20	3/16/78	1978	3	16	0	Alumbramiento de aguas subterráneas en la provincia de Almería. (181/00008	Gómez Angulo, Juan Antonio
21	3/17/78	1978	3	17	0	Decretos 1336/1977 y 320/1978 sobre Cámaras Agrarias, la gestación, democr	Zapatero Gómez, Virgilio
22	3/17/78	1978	3	17	0	Profesores de enseñanza permanente de adultos. (181/000054)	Gutiérrez Pascual, Vicente
23	3/17/78	1978	3	17	0	Propósitos del Gobierno en relación con la empresa privada Minas de Figaredo	Palacio Alvarez, Manuel
24	3/17/78	1978	3	17	0	Pérdidas originadas por varias avenidas en el río Ebro. (181/000083)	Cristóbal Montes, Angel
25	3/17/78	1978	3	17	0	Personal de centralitas dependientes de la Compañía Telefónica Nacional de E	Sáenz Cosculluela, Javier Luis
26	3/23/78	1978	3	23	0	Trasvase Tajo-Segura. (181/000017)	Fuente y de la Fuente, Licinio d
27	3/27/78	1978	3	27	0	Política de producción y comercialización del vino. (181/000011)	Sáenz Cosculluela, Javier Luis
28	3/28/78	1978	3	28	0	Política del Ministerio de Agricultura referente al paro agrícola. (181/000012)	Fuente y de la Fuente, Licinio d
29	3/28/78	1978	3	28	0	Política de orientación de las producciones agrarias. (181/000015)	Fuente y de la Fuente, Licinio d
30	3/28/78	1978	3	28	0	Agricultura de grupo. (181/000016)	Fuente y de la Fuente, Licinio d

GENDER	PARTY	PARLIAMENTA	RESULT	TYPE	COMMITTEE NAME	COMMITTEE (	CODE	SUBCODE	CODE_2	SUBCODE_2	Autor2	Autor3
1	GUCD	10	Convertido	Ordinaria	Comisión de Economía y Hacienda	3	1	107				
1	GUCD	10	Convertido	Ordinaria	Comisión de Economía y Hacienda	3	4	402				
1	GMC	4	Tramitado po	Ordinaria	Comisión de Educación	6	6	601				
1	GS	2	Caducado	Ordinaria	Comisión de Agricultura	12	21	2104				
1	GS	2	Tramitado po	Ordinaria	Comisión de Agricultura	12	4	402				
1	GS	2	Tramitado po	Ordinaria	Comisión de Sanidad y Seguridad Social	10	3	322				
1	GCO	1	Caducado	Ordinaria	Comisión de Trabajo	10	20	2004				
1	GS	2	Convertido	Ordinaria	Comisión de Interior	4	20	2001				
1	GS	2	Tramitado po	Ordinaria	Comisión de Agricultura	12	21	2104				
1	GS	2	Tramitado po	Ordinaria	Comisión de Agricultura	12	21	2103				
1	GS	2	Tramitado po	Ordinaria	Comisión de Agricultura	12	4	402				
1	GS	2	Tramitado po	Ordinaria	Comisión de Agricultura	12	4	402				
1	GS	2	Tramitado po	Ordinaria	Comisión de Agricultura	12	4	402				
1	GSC	2	Caducado	Ordinaria	Comisión de Trabajo	10	5	510				
0	GS	2	Tramitado po	Ordinaria	Comisión de Presidencia	9	20	2004				
1	GS	2	Tramitado po	Ordinaria	Comisión de Educación	6	6	601				
1	GUCD	10	Convertido	Ordinaria	Comisión de Hacienda	3	15	1501				
1	GS	2	Tramitado po	Ordinaria	Comisión de Obras Públicas y Urbanismo	8	10	1002				
1	GS	2	Tramitado po	Ordinaria	Comisión de Interior	4	20	2001				
1	GUCD	10	Decaído	Ordinaria	Comisión de Obras Públicas y Urbanismo	8	21	2104				
1	GS	2	Tramitado po	Ordinaria	Comisión de Agricultura	12	4	402				
1	GS	2	Tramitado po	Ordinaria	Comisión de Educación	6	6	604				
1	GS	2	Tramitado po	Ordinaria	Comisión de Industria y Energía	1	8	805				
1	GS	2	Decaído	Ordinaria	Comisión de Obras Públicas y Urbanismo	8	21	2104				
1	GS	2	Caducado	Ordinaria	Comisión de Trabajo	10	17	1706				
1	GAP	1	Convertido	Ordinaria	Comisión de Agricultura	12	21	2104				
1	GS	2	Tramitado po	Ordinaria	Comisión de Agricultura	12	4	404				
1	GAP	1	Convertido	Ordinaria	Comisión de Agricultura	12	4	402				
1	GAP	1	Convertido	Ordinaria	Comisión de Agricultura	12	4	402				
1	GAP	1	Convertido	Ordinaria	Comisión de Agricultura	12	4	402				

2. Spanish codebook English version adapted:

CODEBOOK for the MEDIA	
Topic	**Codigo**
1. Macroeconomía	
	101. Inflación, precios y tipos de interés
	103. Desempleo
	104. Política monetaria, Banco de España (Central Nacional), reserva monetaria, tasa de descuento
	105. Presupuestos, gasto público y ley de presupuestos
	107. Impuestos, política fiscal y reforma tributaria
	108. Política industrial
	110. Control y estabilización de precios
	199. Otros
2. Derechos, libertades civiles, problemas relativos a las minorías	**200. General**
	201. Minorías étnicas y discriminación racial
	202. Discriminación de género e igualdad de derechos. Discriminación a homosexuales y derechos de las parejas de un mismo sexo
	204. Discriminación relativa a la edad
	205. Discriminación relativa a las personas con enfermedades o discapacitadas
	206. Derechos y cuestiones sobre voto, participación y representación política
	207. Libertad de expresión y religión. Igualdad de derechos en general. ABORTO.
	208. Derecho a la privacidad y al acceso a la información
	209. Actividades contra el Estado
	299. Otros
3. Salud	**300. General**
	301. Reformas generales del Sistema Nacional de Salud (SNS)
	302. Cuestiones generales sobre la cobertura del SNS, seguro, costes y derecho a tratamiento.
	321. Regulación de la industria farmacéutica y otros servicios sanitarios como los dentistas
	322. Instalaciones sanitarias, hospitales, construcción y pago en el sistema sanitario. .
	323. Acuerdos entre la Seguridad Social (o NHS) y compañías privadas. Proveedores, pagos de seguros y regulaciones.
	324. Negligencia médica, malas prácticas, fraude y sistemas de compensación
	325. Recursos humanos, educación y formación. Personal de sanidad
	331. Prevención de enfermedades, tratamiento y promoción de la salud
	332. Menores de edad
	333. Enfermedades mentales
	334. Tratamiento de larga duración, servicios de rehabilitación, enfermos terminales y problemas relativos al envejecimiento
	335. Gasto farmacéutico, consumo público y precios de los medicamentos
	341. Tabaco
	342. Alcohol, Control de drogas ilegales, Temas genericos relacionados con drogas ilegales.

	398. Investigación y desarrollo en salud
	399. Otros
4. Agricultura e industria pesquera	400. Agricultura General
	401. Exportaciones e importaciones agrícolas
	402. Subsidios y regulación en agricultura
	403. Inspección de alimentos y seguridad alimentaria
	404. Promoción y márketing agrícola
	405. Enfermedades animales, de cultivos y controles
	408. Política pesquera y caza
	498. Investigación y desarrollo en agricultura y ganadería
	499. Otros
5. Trabajo	500. General
	501. Entorno y condiciones laborales, accidentes laborales y sistemas de compensación
	502. Política laboral activa, de formación y desarrollo de la mano de obra
	503. Pensiones y jubilación anticipada. Otros beneficios derivados del trabajador
	504. Cuestiones generales a cerca de los sindicatos
	505. Cuestiones generales sobre política de empleo y negociación colectiva. Lew de Empleo y regulación del mercado de trabajo.
	506. Empleo y juventud
	529. Trabajo e inmigración
	599. Otros
6. Educación y cultura	600. General
	601. Formación universitaria
	602. Educación primaria y secundaria
	603. Educación especial para estudiantes con dificultades de tipo social, económico, etc.
	604. Formación profesional
	606. Educación especial para estudiantes con algún tipo de minusvalía.
	607. Excelencia Educativa.
	698. Investigación en educación
	699. Otros
7. Medio Ambiente	700. General
	701. Calidad del agua, polución y conservación de la costa
	703. Eliminación de desperdicios y basuras
	704. Problemas relativos a substancias, fluidos y desperdicios contaminantes y toxicos
	705. Contaminación del aire, ruido y calentamiento global
	707. Reciclaje
	708. Amenazas medioambientales procedentes del entorno interior
	709. Protección de especies y bosques
	711. Conservación de la tierra y el agua. Cuestiones del medio ambiente relacionados con la agricultura
	798. Investigación y desarrollo en medio ambiente

	799. Otros
8. Energía	800. General
	801. Energía Nuclear
	802. Electricidad y Hidroelectricidad
	803. Gas natural y petróleo (incluyendo instalaciones offshore)
	805. Minas y carbón
	806. Energías alternativas y renovables
	807. Conservación de la energía
	898. Investigación y desarrollo:
	899. Otros
900. Inmigración y refugiados	900. Inmigración y refugiados
10. Transporte	1000. General
	1001. Transporte público y seguridad
	1002. Construcción de carreteras, mantenimiento y seguridad en las carreteras
	1003. Aeropuertos, tráfico aéreo y seguridad
	1005. Transporte ferroviario y seguridad
	1007. Asuntos Marítimos e industria naval
	1010. Obras pública y servicios de transporte
	1098. Investigación y desarrollo en temas de transporte
	1099. Otros
12. Crimen y justicia	1200. Cuestiones generales
	1201. Policía y autoridades de lucha contra el crimen, control de armas, fuerzas de seguirdad privada. Agencias que tratan el crimen o la ley.
	1202. Crimen financiero y crimen organizado. Fraude fiscal. Delito de cuello blanco.
	1203. Crímenes relativos al narcotráfico y consumo de drogas
	1204. Sistema judicial y administración de los tribunales
	1205. Carceles
	1206. Crimen juvenil
	1207. Abusos a menores y pornografía infantil
	1208. Violencia doméstica y violencia de género
	1210. Código penal y acciones civiles
	1211. Prevención del crimen
	1227. Terrorismo y lucha contra el terrorismo
	1299. Otros
13. Política social	1300. General
	1302. Pobreza y asistencia a las familias con pocos ingresos
	1303. Políticas orientadas a personas mayores
	1304. Asistencia a los discapacitados y personas con minusvalías
	1305. Asociaciones de voluntarios y fundaciones
	1308. Vida familiar y trabajo. Cuidado del menor

	1399. Otros
14. Planificación urbanística y política de vivienda	1400. General
	1401. Política de vivienda en las ciudades
	1403. Desarrollo económico urbano y problemas generales en las ciudades
	1404. Política de vivienda en zonas rurales
	1405. Desarrollo económico rural y problemas generales de las zonas rurales
	1406. Viviendas de protección oficial
	1408. Residencias y viviendas para personas mayores y con minusvalías.
	1409. Programas de viviendas para personas sin casa, indigentes
	1499. Otros
15. Política industrial y comercio	1500. General
	1501. Política bancaria
	1502. Mercado de valores
	1504. Hipotecas, tarjetas de crédito y otros sistemas de crédito bancario
	1505. Seguros
	1507. Suspensión de pagos, bancarrota e insolvencia
	1520. Legislación antitrust
	1521. Problemas relativos a la pequeña y mediana empresa y pequeño comercio
	1522. Derechos de propiedad y patentes
	1523. Ayuda previstas en caso de desastres naturales, fuegos y accidentes
	1524. Turismo
	1525. Política de protección al consumidor y protección de datos
	1526. Loterías y apuestas
	1599. Otros
16. Política de defensa	1600. General
	1602. Alianzas en política de seguridad y defensa (OTAN)
	1603. Inteligencia militar y espionaje, CIA
	1604.Capacidad de las fuerzas armadas
	1605. Controles a la proliferación de armas.
	1606. Ayuda militar y venta de armas a otros paises
	1608. Recursos humanos de las fuerzas armadas
	1610. Adquisición de armas y compra de material militar
	1611. Instalaciones militares, propiedad y edificios
	1614. Problemas de medio ambiente causados por acciones militares
	1615. Fuerzas armadas y protección civil
	1616. Personal civil y el empleo de la industria de defensa
	1617. Contratos militares
	1619. Participación directa en conflictos bélicos.
	1620. Violaciones de DDHH en tiempo de guerra y denuncias contra las fuerzas armadas
	1698. Investigación y desarrollo en cuestiones militares

	1699. Otros
17. Investigación, tecnología y comunicaciones	1700. General
	1701. Misiones de carácter aeroespacial
	1704. Satélites y otros instrumentos aeroespaciles de uso comercial
	1705. Transferencia de tecnología científica y cooperación internacional
	1706. Servicios de telecomunicaciones y de telefonía.
	1707. Medios de comunicación
	1708. Previsión del tiempo y problemas geológicos
	1709. Industria informática y seguridad informática
	1798. Proyectos de I+D
	1799. Otros
18. Comercio exterior	1800. General
	1802. Acuerdos de libre comercio, conflictos y regulación
	1803. Promoción a la exportación y regulación
	1804. Inversión extranjera. Inversiones en España e inversiones españolas en el exterior
	1806. Competitividad y balanza de pagos
	1807. Importaciones y regulación de importaciones.
	1808. Mercado de divisas y tipo de cambio
	1899. Otros
19 . Política Exterior	1900. General
	1901. Ayuda al desarrollo y cooperación internacional
	1902. Acuerdos internacionales relativos al medio ambiente
	1905. Paises en vías de desarrollo
	1906. Sistema financiero internacional y organizaciones económicas internacionales
	1910. Unión Europea: cuestiones institucionales
	1921. Pais o región especifica
	1925. Derechos humanos
	1926. Organizaciones internacionales
	1927. Terrorismo internacional
	1929. Diplomacia
	1999. Otros
20. Gobierno y Administración Pública	2000. General
	2001. Relaciones intergubernamentales y entidades locales
	2002. Eficacia de la administración pública
	2003. Servicio postal
	2004. Administración Pública , beneficios para empleados del gobierno
	2005. Nombramientos y nominaciones (no codificables en otra parte)
	2006. Premios y reconocimientos
	2007. Contratos públicos, subcontratación de servicios y mal uso de recursos públicos y corrupción

	2008. Privatización del sector público y nacionalizaciones.
	2009. Administración de hacienda
	2011. Parlamento y Constitución
	2012. Regulación de actividades políticas, elecciones y campañas electorales
	2015. Valoraciones, quejas y denuncias contra el gobierno, la Administración Pública o los políticos en general.
	2030. Días festivos y fiestas nacionales
	2099. otros
21. Recursos naturales y gestión del agua	**2100. General**
	2101. Parques naturales y áreas protegidas
	2103. Utilización de recursos naturales
	2104. Recursos hídricos: desarrollo, obras públicas y puertos
	2199. Otros
23. Eventos culturales, arte y humanidades	**2300. General**
	2301. Cine, teatro, música y danza
	2302. Publicación de libros y obras literarias en general
	2399. Otros
27. Inclemencias meteorológicas y desastres naturales	**2700. General**
29. Eventos deportivos	**2900. General**
30. Obituarios y sucesos mortales	**3001. Muerte natural**
	3002. Muerte violenta
	3099. Otros

3. Diputados (first 30 examples):

num	id	name	att	source
1	1	Gervasio Martínez-Villaseñor García	Diputado en las Legislaturas Constituyente, I, IV y V y Senador en la III. Casado. Cuatro hijos. Maestro nacional. Licenciado en Derecho. Abogado. Funcionario. Vicepresidente del Grupo Parlamentario de UCD 1978-82. Pre	leg6
2	2	José Vicente Beviá Pastor	Diputado en las Legislaturas II, III, IV y V y Senador en la Constituyente y I. Casado. Tres hijas. Licenciado en Filosofía y Letras (Sección de Filología Clásica). Catedrático de Griego de INB. Profesor Universitario. Vicepresid	leg6
3	3	Luis Alberto Aguiriano Forniés	Diputado en las Legislaturas III, IV y V y Senador en la Constituyente y II. Casado. Un Hijo. Economista y Técnico en publicidad	leg6
4	4	Presentación Urán González	Diputada en la V Legislatura. Dos hijos. Administrativa	leg6
5	5	Luis Felipe Alcaraz Masats	Diputado en las Legislaturas I y V. Doctor en Filología Románica por Granada. Miembro del Consejo Federal de IU.	leg6
6	6	María Jesús Aramburu del Río	Un hijo. Licenciada en Filosofía y Letras (Filología Hispánica). Profesora. Miembro del Consejo Federal de IU. Miembro de la Permanente, Ejecutiva, Consejo Andaluz de IU-CA, Los Verdes. Responsable de Formación Teór	leg6
7	7	Julio Villarrubia Mediavilla	Casado. Dos hijos. Licenciado en Derecho. Secretario Interventor de Administración Local. Abogado. Secretario General del PSOE de Palencia y miembro del Comité Federal. Concejal, Primer Teniente de Alcalde y Portavo	leg6
8	8	Eugenio Enrique Castillo Jaén	Diputado en la Legislatura V. Casado. Dos hijas. Licenciado en Farmacia y empresario, especialista en alimentación y ecología. Del Comité Ejecutivo Provincial desde 1992 y de la Junta Directiva Nacional del PP desde 199	leg6
9	9	José Madero Jarabo	Diputado en la Legislatura V. Casado. Ingeniero Agrónomo. Funcionario de la Administración Central del Estado. Diputado en las Cortes de Castilla-La Mancha 1987-91.	leg6
10	10	José Luis Rodríguez Zapatero	Diputado en las Legislatura III, IV y V. Casado. Dos hijas. Licenciado en Derecho. Abogado. Profesor de Derecho Político. Secretario General del PSOE de León. Miembro del Comité Federal.	leg6
11	11	María Amparo Valcarce García	Casada. Una hija. Licenciada en Geografía e Historia. Inspectora de Educación. Consejera Comarcal del Bierzo 1991-95. Concejala del Ayuntamiento de Fabero desde 1991.	leg6
12	12	Javier Ignacio García Gómez	Casado. Una hija. Mediador de seguros. Concejal del Ayuntamiento de Jaén. Diputado Provincial de Jaén. Secretario Provincial del PP en Jaén.	leg6
13	13	Luis de Torres Gómez	Diputado en las Legislaturas IV y V y Senador en la III. Casado. Ocho hijos. Maestro Nacional. Concejal. Alcalde de Andújar. Diputado Provincial. Diplomado CESEDEN. Miembro del Comité Ejecutivo Regional y de la Junta D	leg6
14	14	Gabino Puche Rodríguez-Acosta	Diputado en la Legislatura V y Senado en la III y IV. Casado. Dos hijas. Licenciado en Ciencias Económicas y Empresariales. Funcionario del Cuerpo de Intervención y Contabilidad de la Seguridad Social. Miembro del Comit	leg6
15	15	Isidoro Hernández-Sito García-Blanco	Diputado en las Legislaturas I, IV y V. Casado. Una hija. Bachiller Superior. Agricultor- Ganadero. Consejero de Agricultura y Comercio de la Junta Pre-Autonómica 1980-82. Diputado de la Asamblea de Extremadura 1983-89	leg6
16	16	Francisco Zambrano Vázquez	Casado. Dos hijos. Doctor en Medicina y Cirugía. Maestro Nacional. Médico Inspector del Equipo Territorial de Inspección Sanitaria del Ministerio de Sanidad y Consumo en Extremadura.	leg6
17	17	Manuel Núñez Pérez	Diputado en las Legislaturas Constituyente, I, II, III, IV y V. Casado. Cuatro hijos. Licenciado en Derecho. Fundador de UCD. Secretario Organización. En la actualidad miembro del Comité Ejecutivo y de la Junta Directiva del	leg6
18	18	Juan Morano Masa	Diputado en las Legislaturas IV y V. Casado. Cuatro hijos. Licenciado en Derecho. Abogado en ejercicio de los Colegios de León y Valladolid.	leg6
19	19	Angel Escuredo Franco	Diputado en la Legislatura V. Casado. Dos hijos. Maestro Industrial. Alcalde de Villadecanes Toral de los Vados 1983-95. Diputado provincial 1987-91. Procurador de las Cortes de Castilla y León 1993-95.	leg6
20	20	María Visitación Pérez Vega	Diputada en la Legislatura V. Soltera. Diplomada en Derecho.	leg6
21	21	Amador Álvarez Álvarez	Diputado en la Legislatura V. Casado. Dos hijos. Profesor de EGB, especialidad en ciencias. Alcalde de Carrascalejo desde el 19 de abril de 1979. Diputado Provincial de Cáceres desde el año 1987. Portavoz del Grupo Popu	leg6
22	22	Andrés Ollero Tassara	Diputado en las Legislaturas III, IV y V. Soltero. Catedrático de Filosofía del Derecho. Presidente de la Comisión de Justicia y de la de Investigación Científica y Desarrollo Tecnológico del PP. Miembro del Comité Ejecutivo Re	leg6
23	23	Francisco Antonio González Pérez	Diputado en la Legislatura V. Casado. Dos hijos. COU. Ingeniería Industrial (sin concluir). Jefe de Negociado de Compañía Naviera. Presidente Nuevas Generaciones 1983-88. Presidente del PP de Ceuta 1988-93. Concejal d	leg6
24	24	Diego Jordano Salinas	Diputado en las Legislaturas III, IV y V. Casado. Cuatro hijos. Licenciado en Derecho. Especialidad en Derecho de la Empresa. Abogado. Ha sido Presidente del PP de Córdoba 1986- 93. Secretario General del PP de Andalu	leg6
25	25	Manuel Francisco Alcaraz Ramos	Casado. Doctor en Derecho. Profesor de Derecho Constitucional en la Universidad de Alicante. Miembro del Consejo Político de EUPV y del Consejo Federal de IU. Ha sido Concejal de Cultura del Ayuntamiento de Alicante	leg6
26	26	Manuel Arqueros Orozco	Diputado en las Legislaturas IV y V y Senador en la III. Casado. Tres hijos. Licenciado en Derecho. Abogado en ejercicio, perteneciente al Ilustre Colegio de Abogados de Madrid. Afiliado al PP desde 1976.	leg6
27	27	Teresa Cunillera i Mestres	Diputada en la II Legislatura. Funcionaria. Miembro del Consell Nacional y de la XVIII Federación del PSC. Directora del Gabinete del Ministro de Relaciones con las Cortes y del Secretario del Gobierno 1987-93. Asesora en	leg6
28	28	Vicente Martínez-Pujalte López	Diputado en la Legislatura V. Licenciado en Ciencias Económicas. Técnico (excedente) de la Cámara Oficial de Industria, Comercio y Navegación de Valencia. Profesor de la Universitat de Valencia. Miembro del Comité Ejecu	leg6
29	29	María Dolores Calderón Pérez	Diputada en la Legislatura V. Concejala del Ayuntamiento de Morón de la Frontera. Miembro del Comité Ejecutivo Provincial del PP de Sevilla.	leg6
30	30	María José Camilleri Hernández	Casada. Tres hijos. Licenciada en Derecho por la Universidad de Sevilla. Durante diez años Adjunta al Defensor del Pueblo andaluz.	leg6

4. Precision, recall and F1-score of every class and subclass (all the CODE part and the part of SUBCODE 1):

CODE	Precision	Recall	F1
1	0.6503	0.6611	0.6556
2	0.5548	0.5226	0.5382
3	0.8031	0.8093	0.8062
4	0.7638	0.8346	0.7976
5	0.7528	0.784	0.7681
6	0.7894	0.8366	0.8123
7	0.7099	0.6647	0.6866
8	0.874	0.75	0.8073
9	0.5926	0.5517	0.5714
10	0.8459	0.9232	0.8829
12	0.8463	0.8528	0.8496
13	0.6891	0.652	0.67
14	0.7353	0.6667	0.6993
15	0.7083	0.5129	0.595
16	0.8779	0.8367	0.8568
17	0.7802	0.6283	0.6961
18	0.56	0.3784	0.4516
19	0.75	0.8207	0.7838
20	0.6667	0.6526	0.6596
21	0.9056	0.8779	0.8915
23	0.6667	0.4167	0.5128
-	-	-	-
SUBCODE_1	Precision	Recall	F1
0	0.6	0.5	0.5455
1	0.7778	0.5385	0.6364
3	1.0	0.75	0.8571
4	0.7143	0.7143	0.7143
5	0.7797	0.92	0.844
7	0.8429	0.8939	0.8676
8	0.9444	0.7727	0.85
10	0.0	0.0	0.0
99	1.0	0.6667	0.8

5. Class and subclass of political speeches attached with politicians´ characteristics (first 30 examples):

CODE	SUBCODE												
12	1208	mujer	sin hij										
3	343	mujer	solter	hij	bng	secretari	diputad						
3	302	mujer	casad	hij	diputad	licenciad	filosofía						
3	398	hombre	sin hij	funcionari									
12	1203	hombre	casad	sin hij	socialista	secretari	doctor	médico	químicas				
13	1300	hombre	casad	hij	profesor	senador	diputad	doctor					
6	699	mujer	casad	hij	diputad	licenciad	ciencias económicas	empresariales					
12	1205	mujer	sin hij	funcionari									
6	699	hombre	hij	delegad	diputad	médico							
16	1699	hombre	sin hij	profesor	licenciad	doctor	historia	filosofía					
10	1002	hombre	casad	hij	funcionari	alcalde	profesor	delegad	técnic	diplomad	ciencias políticas		
2	207	mujer	sin hij	psoe	secretari	diputad	diplomad	licenciad	doctor				
13	1304	hombre	sin hij	autor	secretari								
6	600	hombre	sin hij										
10	1002	hombre	solter	sin hij	autor	catedrático	profesor	licenciad	doctor	máster	ciencias políticas	derecho	sociología
16	1602	hombre	casad	hij	alcalde	secretari							
21	2101	hombre	casad	hij	pp	profesor	senador	diputad	técnic				
12	1208	mujer	casad	hij	autor	profesor	delegad	portavoz	licenciad				
10	1007	hombre	casad	sin hij	popular	directiv	diputad	licenciad	derecho				
10	1002	mujer	solter	sin hij	pp	secretari	directiv	diputad	estudiante				
3	399	hombre	sin hij	psc	funcionari	secretari	ministr	diputad					
10	1003	mujer	sin hij	secretari	doctor								
10	1002	hombre	sin hij	senador	diplomad								
10	1005	hombre	solter	sin hij	diputad	licenciad	abogad	derecho					
9	900	hombre	casad	hij	profesor	portavoz	diputad	licenciad	filosofía				
17	1706	hombre	sin hij	cdc	secretari	diputad	licenciad	derecho					
16	1611	hombre	casad	hij	izquierda unida	secretari	licenciad	médico					
5	505	hombre	sin hij	técnic									
3	300	hombre	solter	sin hij	cdc	secretari	diputad	licenciad	derecho				
10	1000	hombre	casad	hij	psoe	secretari	senador	licenciad	médico				

6. Correlation found with logistic regression on CODE level (some features with all classes):

CODE	hombre	mujer	casad	solter	divorciad	separad	viud	hij	sin hij	a la valencian
1	-0.813800977275163	-0.8974158977400103	0.2783748478576264	0.0665811302990275	0.14428329882493726	0.7540587622670896	-0.156544329634715 57	-0.9930835697466646	-0.7181333052686537	-0.071643302
2	-1.2591857916912206	-0.4634523235613372	-0.19291258276109377	-0.3534622833940992	0.12970890615733882	-0.045728983902131946	0.9076911445789057	-0.7879939382220057	-0.9346441770308366	-0.072656374
3	-0.8672888665652448	-0.4365151167838136	-0.1751347075406698	-0.14594473196486143	0.07004875887578131	0.1440461011856234	-0.5171089275305382	-0.5490068627063879	-0.7547971206427764	1.2962882734
4	-0.6449046934177847	-0.7093967803888526	-0.03551671458448561	-0.03209762373415343	0.3878622220186283	-0.04907741153575245	0.9770185477976195	-0.7554001990955556	-0.598901274711076	-0.070718882
5	-0.6823429802485763	-0.7502688544984706	0.10521445257767481	0.08413769591250944	-0.518325252747333	-0.4954052178573823	-0.368424012888855 07	-0.6557014459861082	-0.7769103887611702	-0.174707298
6	-0.7889453812114737	-0.528602764604017	-0.18805563851110826	-0.043257527456921956	-0.007963957939352 17	0.014187335452211378	-0.6628193774773131	-0.5878065914346913	-0.7297415543812862	0.4978941158
7	-0.7138111117484803	-0.7883990376109018	0.25039987938993696	0.12652176687564443	0.04288300399614517	0.6544051316052555	0.2083384724352265	-0.9304101081028889	-0.5718000412567621	-0.186963436
8	-0.8627687898587807	-0.8927304825012107	-0.3020286581246194	-0.010538533594903237	0.1611614625215056	-0.13147244017165735	-0.4829295961755669	-0.9675179656663756	-0.7879813066936318	-0.060194725
9	-1.4042837924382507	-1.3321962447096147	0.5115443794161145	-0.17208224664427052	-0.3261124262159964	-0.32255366950390013	-0.06994270128144452	-1.4057554963538474	-1.330724540794021	-0.019216505
10	-0.4545744158114517	-0.7045932164600435	0.16860884384593627	-0.03681554816772055 6	-0.19122685492791508	-0.1128516226726 4963	-0.1098677798663249	-0.641505653280647	-0.5176619789905633	-0.213178809
12	-0.6228802255180179	-0.3903467710591456	-0.07326765422063614	0.193345637184452 5	-0.3266797479015813	0.1549220452922 2303	-0.7714418709118733	-0.4637569241556572	-0.5494700724216703	-0.273772743
13	-1.0164195749070388	-0.560787757954093	-0.261116638306704	-0.32574869409135293	-0.5940291125329229	0.213776724266788 2	-0.3162664751691222	-0.7954478313586532	-0.7817595015024649	-0.050903680
14	-0.8760228035885	-0.9473948161761591	-0.5874476635771496	0.33426026299618167	-0.2743986165824921	-0.5128722099596187	-0.3112866371021355	-0.6151992621355654	-1.2082183576290753	-0.094201955
15	-0.6130508399803745	-0.6895068338449729	-0.3424481099388927	0.11268324855261441	0.15586388664290243	-0.6407572860065713	0.2080225544137544 7	-0.4782573746334866	-0.8243002991919572	-0.042315005
16	-0.2540114255709481	-1.1538797598672623	-0.0732822463576377	-0.35476251280688825	0.057885161720100234	-0.12388513440280009	0.5573216918403737	-0.6448373887739043	-0.7630537966646725	-0.144078252
17	-0.8322600879560911	-0.8977070212059126	0.03714536759820966	-0.24053975245828554	0.5541170242272114	0.5106279239932761	-0.1825152704335 1143	-1.032808942823846	-0.697158166338087	-0.031410982
18	-0.8920963850100979	-1.382335297532789	-0.5021243721027159	0.03664372898844322	0.08346754610525754	-0.565351303342615	-0.0933856221618049	-0.9775328919178495	-1.2968987906249445	-0.008656979
19	-0.46266214706689174	-0.7067547368146583	-0.02428365094777086	-0.03108703420750722 6	-0.29118124625215175	-0.2202601471199116	0.2749594356072639	-0.5976626267144576	-0.5717542571670625	-0.280095666
20	-0.6581723042196849	-0.7838317251504444	0.22128723148865542	0.29960486404470427	0.3192584739851653	-0.7649652721939441	0.1035424353288375 3	-0.7986954313220259	-0.6433085980479225	-0.169512717
21	-0.8652219023220886	-0.6770132051557518	-0.2070518876274846 4	-0.27863602763567386	-0.08895452866605692	-0.018716865036618152	-0.3051638362902079 3	-0.7176367720789641	-0.824598335398524	-0.089951776
23	-0.9242754410460844	-0.9606101696914562	0.14301685212089008	-0.6021266935357849	-0.2399821631706264	0.4576758118364222	-0.1361702311040493	-0.7785077630999574	-1.1063778476374377	-0.054443680

7. Correlation found with logistic regression on SUBCODE level (some features with some subclasses):

SUBCODE	hombre	mujer	casad	solter	divorciad	separad	viud	hij	sin hij	a la v
100.0	-1.2425697003237859	-1.231651042027942	-0.26310101323453282	-0.4036769295980153	-0.2818880182742841	0.43413292027784983	-0.024925731562888598	-0.9251495482712248	-1.5490711940805175	-0.00
101.0	-1.3680443897558574	-1.284067657030328	-0.9597940306856444	-0.2992431211649828 6	-0.20448139102831497	-0.17869432378736155	-0.029387869721253 23	-1.5356055513758586	-1.1165064954103343	-0.02
103.0	-1.597951221639904	-1.2715207880556945	-0.5114321093166025	-0.11726725509670208	0.3611056895019163	-0.226467497373328	-0.05741762985643806	-1.2216886352248818	-1.6477833744467998	-0.00
104.0	-1.2761888951810737	-1.478109832456448	-1.236182557505706	-0.5704338276189126	-0.11368974759688709	-0.11927306820024403	-0.027701208027595275	-1.627613620218609	-1.1266846074189103	-0.00
105.0	-1.0090085691150918	-1.2661327552299357	0.5969353243086172	-0.5837045984613415	0.27444287229322023	-0.06339394093494703	-0.07811101217045271	-1.2102164918153695	-1.0649248325989635	-0.03
107.0	-1.0485159059079485	-1.0648929515154915	-0.1537101578197574	-4.976663268205759e-05	-0.6507966095328206	0.8329070893988928	-0.08176602733454104	-1.119637051404738	-0.9937718060188464	-0.04
108.0	-1.2495404504973042	-1.173972902694591	0.09284385363358931	0.5087310630904214	-0.27558040828362257	-0.20347105011832353	-0.022452226154986 82	-1.4670624455269201	-0.956450907664987	-0.00
152.0	-1.4106582290059695	-1.3572405248276695	-0.7792232533888558	-0.3317010175296733	-0.08547034884730134	-0.11618258352422056	-0.008355586522664375	-1.2131724719879518	-1.5547262818456884	-0.00
199.0	-1.635752202100226	-0.9205620202872404	-0.8636430822040353	-0.6073553985092137	0.48129818583336476	-0.207076882639369 08	-0.016056506764183548	-1.042902485764438	-1.5134117366230249	-0.00
200.0	-1.383214763443417	-1.0839256493873977	-0.5236986187022157	-0.2517203612330608	-0.32552061363182405	-0.29594483449365755	0.7510688745497598	-0.9263480306156628	-1.5407923822151388	-0.00
201.0	-1.6971837116457114	-0.9278884206704509	-0.5395995598502692	-0.30836033411454383	-0.2740701907459837	-0.22467657226260973	-0.04290492456767193	-1.275607990342631	-1.3494641419735522	-0.00
202.0	-2.006644976371049	-0.33146476643065814	-0.2722653538622861	-0.2596658742903332	-0.29873305513761833	0.14228391468507182	0.8690464108723066	-1.002936100214715	-1.3351736425869247	-0.06
205.0	-1.376313877068449	-1.414298750055089	-0.7314879814685696	-0.4522681323317552	-0.19821424420061046	-0.13090739248067 14	-0.02993390108909949	-1.9049964229707277	-0.8856162041527967	-0.02
206.0	-1.5643425543787652	-1.1474669399299473	-0.8923944456879882	-0.5726733546970239	-0.13495693684074414	-0.12111186506401159	-0.017782676285058256	-1.3345355834805028	-1.3772739108282128	-0.00
207.0	-1.560081815266557	-1.1099078853315907	-0.5719891846713102	-0.7343510689540909	-0.2928845951117341	-0.2460310032735248	-0.10826974621268436	-1.3357118606172953	-1.3342778399808237	-0.01
208.0	-1.2753771422185225	-0.9845431605433315	-1.001541806084291	-0.1392519356502631	0.32306167543639813	0.42170727108499356	-0.1321517348437597	-1.1157893660681681	-1.1441309366936918	-0.00
209.0	-1.6067514806827314	-1.199053147881398	-0.8332797973872199	0.20340564508548056	-0.09552857892603298	-1.144567296093808 47	-0.03166451538151827	-1.377177186435842 2	-1.4286274421282839	-0.00
212.0	-1.7099763642817878	-1.035881373705082	-0.21953857333695787	-0.26232712038395245	-0.2349682864579917	-0.20398731292368302	-0.040682459247930354	-1.0852740569419237	-1.6605836810449317	-0.00
230.0	-1.121204623060335	-1.078365246039017	0.09761000510501186	-0.6727280106967152	0.4943769833911164	-0.5549801962202936	-0.07856899873444183	-0.8939325490886249	-1.3056373200106253	-0.00
299.0	-1.6866846108402305	-1.018192391105067	-0.8960846415319941	0.26236218352371676	-0.15141563606014989	-0.21166868325611354	-0.019787622836745 06	-0.8582391660638911	-1.8466378358814093	-0.00
300.0	-1.4190749205544906	-1.1401095006366466	-0.6783223460383927	0.08401704171263667	0.07909448703013962	-0.47062954515547906	-0.07240081753912106	-0.8068072595375346	-1.7523771616535764	-0.02
301.0	-1.4032453921948054	-1.1345177508235365	-0.6777274170675577	-0.4151155911679429	-0.3737930046675378	0.3889536480922031	-0.0816406476737586	-0.9356908221270291	-1.6020723208913166	-0.00
302.0	-1.5146078658392492	-1.0363348312466143	0.02007030020334678 3	0.3077645052884992	-0.5198329797889414	0.3711842918247561	-0.058131488181338832	-1.4102881370704712	-1.1405545600153972	-0.01
321.0	-1.1901582025627573	-1.364431840231251	-0.5498213682433283	0.5263017964763185	0.4096961035294086	-0.29922377111211224	-0.04613905200662152	-1.1229040551429768	-1.4316859876510577	-0.00
322.0	-1.1988379373967901	-0.89434778106438 24	-0.4168533776992958	-0.8847392665179171	-0.5920390186776431	0.38017271355655374	-0.19198720577227005	-0.9046490818330659	-1.1885366366281438	-0.02
323.0	-1.5743911381093392	-1.313472030743071	0.2630945928516422	-0.467855297867087	-0.10078053095148228	-0.11061115912798945	-0.007189202555454	-1.748893543436824	-1.1389696254155925	-0.03
324.0	-1.3167067863082071	-1.4441291060632773	-0.831762045228147	-0.6428752432626446	-0.1293661375490528	-0.15773698066675224	-0.020339259746196714	-1.3873918817850 47	-1.373444010586445	-0.01
325.0	-1.2994249096174952	-0.9532126182923751	-0.9697455084646488	0.10393169762057136	0.01609049287377743	-0.12446449572458544	-0.11221199121102868	-0.9109781160080694	-1.3416594119017955	-0.04
326.0	-1.9173255079506708	-0.9286732440264491	-0.3780964322900408 6	-0.4462232682223925	-0.07976650898818735	-0.09896759901428294	-0.02209673307991883	-1.5108535678239317	-1.3351451841531896	-0.00
327.0	-1.2496835262490298	-1.2055349816039576	-0.6978423132478702	-0.8507249056449002	0.4991042705299338	-0.30172798185881056	-0.029217277458493823	-1.0140377151264122	-1.4411807927265732	-0.00
331.0	-1.6480016699217495	-0.6596294570101336	-1.0007167242058088	-0.5199601247149369	-0.2328467900502647	-0.6304802416177332	0.7335918105379188	-0.47578580025894784	-1.8318453266729327	0.933
332.0	-1.0092275406547708	-1.3813471774060666	-0.9674291037351 01	0.3174720577476307	-0.2538276853087915	-0.2608586573195091	-0.015411405678529506	-1.156244961332599	-1.2343297567282407	-0.00
333.0	-1.4225564807812887	-1.39037284833348679	-1.1016952096089119	-0.409933267525675	-0.0956881565594899	-0.09960825335868169	-0.00745161894570198	-1.5034368197849337	-1.309492509331 22	-0.00
334.0	-1.2585351795520952	-1.1709660068945962	-0.1543905517856135	-0.29702404732478355	-0.23475051360271038	-0.26525732899948906	-0.01686535631021492	-1.2549972770028606	-1.174503909443806	-0.00
335.0	-1.5256932009549289	-1.2623651921192458	-0.5508155023307457	0.08638129435261388	0.9604052803962347	-0.31280362064235695	-0.10289551886371553	-1.2722409443257445	-1.5158174487484255	-0.00
336.0	-1.428143780706 06	-1.3866974889725932	-0.761138466149003	-0.39283012078393704	-0.09075589636319469	-0.13103797955424035	-0.016479719973940453	-1.183335224769972	-1.6315060449086827	-0.00
337.0	-1.3974879031992586	-1.0270215569532601	-0.42877833245068364	0.08150082733670659	-0.20391192329519842	-0.2655557157677028	-0.02857336119670831	-0.8990672484491159	-1.525442211703411	-0.00

8. Correlation found with normalized pointwise mutual information on CODE level (some features with all classes):

CODE	hombre	mujer	casad	solter	divorciad	separad	viud	hij	sin hij
1	0.011367097351433614	-0.018127450304560505	-0.0015758487300410325	0.00014854844718350052	-0.021475808699259698	0.12216620645471629		-0.006444721012736667	0.010237509642084966
2	-0.08641731479634239	0.10856578722505628	0.006683620265696337	-0.026187104750212215	-0.019748544541511107	-0.0016404329118623886	0.196862373404403 88	0.009096447992933115	-0.01496484201205037
3	-0.05249932498135715	0.07140885114025423	-0.011137249125500263	0.017825935950009326	0.019218077000225563	0.03924193656404607	-0.004620354102144 61	0.009505693049170436	-0.015279763878517 9
4	-0.0009116276927643402	0.0013727911736168548	-0.016028106253640893	0.035118967059880414	0.07376603469966674	-0.043928449109706544	0.15618939928954997	-0.01567377930259107	0.023964173786040 02
5	0.005036665581495725	-0.007661590196667953	0.01877188296504657	0.004428759166239373	-0.10634018593434605	-0.088923649955551 8		0.012951885985808676	-0.021057141205968 9
6	-0.02679516106528821	0.03803116117722106	0.002697474753696377	-0.016424333104475858	0.014903417167113185	0.035633303222300634	0.050204767332162295	0.010018510360487811	-0.01584991700473 20
7	0.004148731605687427	-0.006356544672242474	-0.005235284904387359	0.0009837932454770397	0.03645236344111932	0.0560762703399684	0.02449649903169043	-0.005891476086647515	0.009248631326776 3
8	0.010220403666611278	-0.016315561076073373	-0.0399911747109155425	0.028541563473711578	0.09289712684614426	0.00974597214331137		-0.03267856628744751	0.048173051189164 95
9	0.005549955149323096	-0.009108088439289 25	0.05347191109350161	-0.06882298671514933				0.024168264116434118	-0.04685196301839 94
10	0.028265864671939447	-0.041952016829968866	0.01666373668484828 4	-0.012036906365289328	-0.051144599457662222	-0.04470458046590267	-0.09303971118554083	-0.0012387782693505402	0.001856881700426 9
12	-0.0236733261575687 65	0.033445277724180614	-0.0023857761020538667	0.021920015090803234	-0.05475394582995845	0.04891461125180279		-0.0005892155985619174	0.000894876758221 6
13	-0.040215577533967116	0.055631225491344 29	-0.012108529022424219	-0.02525759445439644	-0.09020810225010964	0.04719631963643165		-0.012188239289670 96	0.018980032964781 64
14	0.0028475967533196553	-0.00446288970797427	-0.018121407544801127	0.04577593667801118	0.05427616713309551	-0.021255407674947532		0.008962488539282812	-0.015119243005255 1
15	0.0060989923595315	-0.009408593491962401	-0.021552560765422234	0.03081421308484304 6	0.06438838696104107	-0.10496767025447927	0.02966750287161662	-0.009371252382561703	0.014608426169317 76
16	0.08138409298706584	-0.15057096688473035	0.011167203206333456	-0.06691679036144742	-0.023200312582998216	0.01894670557653317	0.08268476575149047	0.011884154955022385	-0.01920917433889 01
17	0.006221348524599329	-0.009837939236017739	-0.00934145041680018	-0.014572042717600781	0.07780110554635998	0.022126451801682788		-0.02492189770620951	0.037557338615806 9
18	0.039206040333943229	-0.07800536319542815	-0.029097025073904464	0.08200288334440761	0.013166456952144292			-0.012487208153549985	0.019747575700038 17
19	0.024898575906203 27	-0.038258850127185 33	0.015567290790751948	-0.028210889965828568	-0.05550374592896569	-0.035879839030116614	0.036510765496107 91	0.005613253664719281	-0.008744753028681 5
20	0.009680785530728021	-0.014749521358222484	0.000674779420335913 1	0.026268163865165013	0.05617844831436275	-0.10858253519381945	-0.015532721238156 82	-0.005769756427560385 6	0.008937580938408 3
21	-0.013555835308680 34	0.019994369542836888	-0.015385105893604869	0.016390864274635538	0.037675715176379895	-0.0022899745275282617	0.043976550740299 31	7.918884314341845e-05	-0.000127022764126
23	-0.006185963727108 94	0.009460331799302366	0.03569598001985996	-0.11312945794696982	-0.08047210408480748	0.07105502832758338		0.028295057466458 7	-0.052640333549776 3

9. Correlation found with normalized pointwise mutual information on SUBCODE level (some features with some subclasses):

SUBCODE	hombre	mujer	casad	solter	divorciad	separad	viud	hij	sin hij
100.0	0.007093116474329248	-0.01198442193427288	0.03149803135670385	-0.045436400924789 09		0.14486417533930965		0.03318524195789759	-0.07338219950362
101.0	0.006196027540765939	-0.010601064412954764	-0.07459218735997208	0.060723451696378654				-0.08459231647411264	0.089885102727364
103.0	-0.019572152252033678	0.028092688014177265	0.014771273777401594	0.013746968553580334	0.18884784321318376			0.026295883395214346	-0.05636117066894
104.0	0.056733425507647935								0.133114841570620
105.0	0.021184469730181975	-0.037461638463226206	0.03823726311067768	-0.10048229878977595	0.002381162231421034	0.017580863835916363		0.016684616752407448	-0.03050230277277
107.0	0.007323178970819851	-0.011950142519305448	-0.026371135995642243	0.004400911609060677		0.2122401616480187		-0.013255008348853034	0.020858730579453
108.0	-0.0002720476803841266	0.0004363975607934168	-0.016211776545270343	0.11881192902468832				-0.04267955519675563	0.058320291477923
152.0	0.05217734404309665		0.05868138613077984					0.04948433566899236	
199.0	-0.058696215480043425	0.0668604327629364	-0.0005751962282906 85		0.2569477852382076			0.02815308063758941	-0.06404187014540
200.0	-0.0194294310390512	0.02808861263612596	0.005765724622095702	0.029785049322212768			0.34389923519023 85	0.022684906158762636	-0.04549614869316
201.0	-0.07746728426011029	0.08413576171790794	0.008153344399077589	-0.009502389662753791				-0.026175224127801648	0.038198202665509
202.0	-0.1760077289331726	0.164214462269531	0.010996533051695265	-0.0287197009618501	-0.055247039426712594	0.04377491513459092	0.26831978202757095	0.02002551314773766	-0.03631208819377
205.0	0.014012783356749091	-0.025158924055982988	-0.09259455304721009	0.001540082218009236				-0.14894352762083854	0.127222227286943
206.0	0.005654995666682499	-0.009749726552703994	-0.0686019253726 1346			0.33133085632107234		0.0027268360835208264	-0.00485850066703
207.0	-0.0302590435664826	0.04153640446313785	-0.02138860289016494	-0.07282399414999614				-0.03393186282276569	0.048097991369000
208.0	-0.016676473096535614	0.024440794537572814	-0.07450036735329496	0.046606654272542035	0.13765786798642382	0.15285756959091915		-0.03979405271953354	0.054377341633295
209.0	-0.02812948431749833	0.037226756590094306	-0.021625442229815148	0.18800676319977194				-0.030822492691602622	0.042117982475766
212.0	-0.048150401812102346	0.0597011041526374	0.04753211890962627	-0.009502389662753791				0.03519426183796581	-0.08164064314223
230.0	-0.0014388628661731373	0.002266480830252479	0.03926359751031088	-0.11121004589239565	0.079180488754584 7			0.034619335602619	-0.07014659196172
299.0	-0.07510596746029666	0.07671808814467869	0.012726965197661368	0.14103028005697363				0.056701488949680814	
300.0	-0.016830701652965258	0.024747377462047547	-0.007353044798378619	0.061294591018128106	0.06593706055589019			0.029233715585641355	-0.05989634575670
301.0	-0.013259119820867585	0.019794269431154976	-0.0048619341578843 97	-0.019583376664216334		0.1707171995998824		0.02167060256458407	-0.04375858327015
302.0	-0.03444900185249751	0.046745663575016826	0.00756626326364045	0.056085393988522206		0.07676423331428579		-0.01640944843449009	0.025465611355467
321.0	0.020132077867240805	-0.03781190987178701	-0.014568793006148455	0.13495894302598463	0.16559848499684962			-0.0037207085726436 56	0.006221790926488
322.0	-0.018747291722809756	0.027387611373431504	0.006741838545254718	-0.15907357450318071		0.12127286964923055		0.006385349344737067	-0.01108630013497
323.0	-0.008027834837605496	0.012366876997952757	0.07213150925594032					-0.011113611023166247	0.017685200582929
324.0	0.03136138679561532	-0.06893309603107341	-0.051624245949064555					-0.06162437506320508	0.072465030394840
325.0	-0.02160541914941837	0.03103946311114665 4	-0.05045250748048178	0.06670812167292144	0.07048110425644487	0.08568080586094022		-0.0032731323698666904	0.005443909549624
326.0		0.1277965179856441	0.0638053950386268					-0.030822492691602622	0.042117982475766
327.0	0.00656334702776875	-0.011171697162683368	-0.011577344988906224		0.18884784321318376			0.003164842655748392	-0.00556709953077
331.0	-0.08578397251892844	0.09517668666793644	-0.002859919050168394	-0.036131087640 35001	0.06156453170979045		0.28327661412243865	0.04012794211587974	-0.08968095219517
332.0	0.03817162055310351	-0.09005655491329896	-0.07459218735997208	0.15334492459651428				-0.03531786786923859	0.048260773548768
333.0	0.05217734404309665								0.122424810836360
334.0	-0.0002438412639055949 5	0.0003946459832365033	0.008153344399077589	-0.009502389662753791				-0.0037207085726436 56	0.006221790926488

10. Strong correlation found with normalized pointwise mutual information (all classes):

CODE								
1	separad:0.12216620645471628	socialista:-0.11200275511162346	autor:-0.12182267860407535	parlamentario regional:0.2029697	ciencias políticas:0.11124877147	sociología:0.14183226690732542	banca:0.16963944054404698	
2	mujer:0.10856578722505628	viud:0.19686237340440388	psc:-0.10840354064596756	portavoz de economía:0.171805078	farmacia:0.18968463241978664	ciencias económicas:-0.135811782	veterinaria:0.10484301658548448	
3	a la valenciana:0.27177136529033	estudiante:-0.10817654067907373	física:-0.13402956493964646					
4	viud:0.15618939928954995	autor:0.10677483874385024	historia:-0.14001696273801573					
5	divorciad:-0.10634018593434603	cdc:-0.10646340008736987	psc:-0.1095651834648812	convergència:0.1040971464064327	portavoz de economía:0.128907227	empresari:-0.1540588577330806	estudiante:-0.10193227059041014	
6	a la valenciana:0.14012594798473	economista:0.1222928093625329						
7	autor:-0.17301265383083528	parlamento europeo:-0.1242673078	ingenier:-0.1177779843407939	física:-0.1049127118058114				
8	psn:0.2619943364860 42	podemos:0.15183532105095393	autor:-0.1097746110226 7533	directiv:-0.1161379967286862	ministr:0.1310179193977406	economista:0.14921436697635596	maestr:-0.15237856531937385	
9	psc:0.11850727821139663	directiv:0.16556269094937615	conseller:0.15989054794721505	banca:0.3471630640635009	física:0.12380088694040242			
10	izquierda unida:-0.1089120524965							
12	a la valenciana:0.10058983530832	psn:0.1005898353083238	farmacia:0.13095049066887107					
13	izquierda unida:-0.1005517566336	banca:0.15239444467167473						
14	psc:0.11322505176699668	portavoz de economía:0.233940410	médico:-0.11064359504312052	químicas:-0.13182417451286532	sociología:-0.11786949242940975			
15	separad:-0.10496767025447927	portavoz de economía:0.156950814	portavoz de hacienda:0.364540954	geografía:-0.1348050905994727	economista:-0.14741824316620436	banca:0.12362046915721325		
16	mujer:-0.15057096688473035	cdc:-0.1841066203373275	ciudadanos:0.14654854681273896	parlamentario regional:0.2210823	estudiante:-0.10781840047727456	economista:-0.10986395756038933	empresariales:-0.122533860313495	
17	catedrático:-0.1305716145336776	geografía:-0.20489551859716276	farmacia:0.18765159119239785	ciencias políticas:-0.1259356097	veterinaria:0.1071435518800822	maestr:0.10828975642970644		
18	profesor:-0.196433748196039 2	técnic:0.19725660729069527	economista:0.2120689420402717	abogad:-0.1151240923320994				
19	psn:0.12753166800369464	economista:-0.11060927612512356						
20	separad:-0.10858253519381944	pnv:0.3193407301292247	parlamentario regional:0.1117505	empresari:0.16385476266814422	economista:-0.12211859604783296	maestr:-0.1906494529303012		
21	psn:0.21823634538829093	ciudadanos:0.12406949732618006	izquierda unida:-0.1150166842776	empresari:-0.11170622297750027	bachiller:0.11160257782470287	banca:0.1379295170276959		
23	solter:-0.11312945794696982	podemos:0.1128405161116344	bachiller:-0.11116100661591903	geografía:0.1037121279338451	médico:-0.18926094419344114	veterinaria:0.21566731099742326	física:0.10018297263933512	

11. Strong correlation found with normalized pointwise mutual information (some subclasses):

SUBCODE						
100.0	izquierda unida:0.19238332112771	conseller:0.16490987516569985	maestr:0.2082980316221641			
101.0						
103.0	divorciad:0.18884784321318376	empresari:0.2548771519062324	maestr:0.17044681846660764	periodismo:0.3151853677253875	física:0.2811507147883301	
104.0	popular:0.19535844713435485	catedrático:0.24504518867183975				
105.0	parlamentario regional:0.3513765	senador:-0.17703791148992176	banca:0.31804621092966584			
107.0	separad:0.2122401616480187	podemos:0.1643106807446638	químicas:0.17294380819099134			
108.0	ciencias políticas:0.20604249214	sociología:0.17699813465116807	maestr:0.22249367515541305			
152.0	alcalde:0.22323819823162594	profesor:0.16928240401179948	conseller:0.4786598676508558	ingenier:0.4105599256258319		
199.0	divorciad:0.2569477852382076	alcalde:0.18798177107838634	ciencias económicas:0.2364819534			
200.0	viud:0.3438992351902385	izquierda unida:0.18490600017049	farmacia:0.3376351008472571	economista:0.16681348915241764	física:0.2144900243527367	
201.0	cdc:0.20746616097251566	empresari:0.23162779368989825	veterinaria:0.4211301979323113			
202.0	hombre:-0.1760077289331726	mujer:0.16421446226953101	viud:0.2683197820275709			
205.0	psc:0.17382644771235425	portavoz de economía:0.525636350	conseller:0.2118863583084981	economista:0.22126729325243985	empresariales:0.16839459688857135	ciencias biológicas:0.2127815976
206.0	separad:0.33133085632107234	popular:0.24865770389977	autor:0.3403340842666994	directiv:0.2629905443695276	ciencias económicas:0.1963673788	químicas:0.2993619634678631
207.0	cdc:0.2440502895609969					
208.0						
209.0	solter:0.18800676319977194	funcionari:0.17836307672134294	máster:0.18821646147063906	médico:0.2543187977507417		
212.0	sociología:0.16278397583159124	ciencias biológicas:0.2007749065				
230.0	popular:-0.1752342642509532	farmacia:0.207962874543734				
299.0	izquierda unida:0.37885000210948	geografía:0.23440852548275856	médico:0.20734231460794336			
300.0	cdc:0.20453730732482425	bng:0.1892065470921792	ministr:0.1918587483867422			
301.0	separad:0.1707171995998824	bng:0.1832855784690486	parlamento europeo:0.22846577349	empresari:0.2215468066884357		
302.0	ciudadanos:0.22029427272688848	químicas:0.1971758466781501				
321.0	divorciad:0.16559848499684962	psc:0.16278397583159124	ciencias políticas:0.18805103634	economista:0.21022482137167686	sociología:0.16278397583159124	
322.0	podemos:0.20933992035286456	senador:-0.1633917735649201				
323.0	popular:0.1843064927530441	técnic:0.18001147760318145	geografía:0.277844820691105	historia:0.19626464802771307	economista:0.3015741216130348	
324.0	diplomad:0.23216325744684554	doctor:0.2058949101622568	historia:0.19626464802771307	ciencias políticas:0.18399172592	ciencias biológicas:0.3001007502	
325.0	estudiante:0.188637000092164					
326.0	técnic:0.1914083510289384	abogad:0.22017558396919315				